Targeted modulation of protein liquid–liquid phase separation by evolution of amino-acid sequence

Simon M. Lichtinger,[1, a)] Adiran Garaizar,[2] Rosana Collepardo-Guevara,[1, 2, 3, b)] and Aleks Reinhardt[1, c)]
[1)]*Yusuf Hamied Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, United Kingdom*
[2)]*Department of Physics, University of Cambridge, Cavendish Laboratory, Maxwell Centre, JJ Thomson Avenue, Cambridge, CB3 0HE, United Kingdom*
[3)]*Department of Genetics, University of Cambridge, 20 Downing Place, Cambridge, CB2 3EJ, United Kingdom*

Rationally and efficiently modifying the amino-acid sequence of proteins to control their ability to undergo liquid–liquid phase separation (LLPS) on demand is not only highly desirable, but can also help to elucidate which protein features are important for LLPS. Here, we propose a computational method that couples a genetic algorithm to a sequence-dependent coarse-grained protein model to evolve the amino-acid sequences of phase-separating intrinsically disordered protein regions (IDRs), and purposely enhance or inhibit their capacity to phase-separate. We validate the predicted critical solution temperatures of the mutated sequences with ABSINTH, a more accurate all-atom model. We apply the algorithm to the phase-separating IDRs of three naturally occurring proteins, namely FUS, hnRNPA1 and LAF1, as prototypes of regions that exist in cells and undergo homotypic LLPS driven by different types of intermolecular interaction, and we find that the evolution of amino-acid sequences towards enhanced LLPS is driven in these three cases, among other factors, by an increase in the average size of the amino acids. However, the direction of change in the molecular driving forces that enhance LLPS (such as hydrophobicity, aromaticity and charge) depends on the initial amino-acid sequence. Finally, we show that the evolution of amino-acid sequences to modulate LLPS is strongly coupled to the make-up of the medium (e.g. the presence or absence of RNA), which may have significant implications for our understanding of phase separation within the many-component mixtures of biological systems.

Author summary
Protein condensates formed by the process of liquid–liquid phase separation (LLPS) play diverse roles inside cells – from spatio-temporal compartmentalisation to speeding up chemical reactions. When things go wrong, LLPS can have pathological implications. This realisation has boosted the interest in devising approaches to design rationally amino-acid sequence variations to modulate or even reverse the phase behaviour of proteins on demand. Here, we develop an efficient computational method that combines a genetic algorithm with a sequence-dependent coarse-grained model, and an all-atom model for validation, to identify amino-acid sequence variations of intrinsically disordered proteins that intentionally promote or inhibit their LLPS. Our method can be applied to proteins in pure form and within multi-component systems.

I. INTRODUCTION

Liquid–liquid phase separation (LLPS) of multivalent biomolecules (e.g. proteins and nucleic acids) is an important mechanism employed by cells to control the spatio-temporal organisation of their many components.[1,2] Biomolecular condensates, or membraneless organelles, such as stress granules,[3] P-granules,[4,5] the nephrin–NCK–WASP system[6] and the nucleoli,[7] are formed by LLPS and have diverse biological functions. LLPS inside cells plays a very diverse range of roles beyond membraneless compartmentalisation, such as in gene silencing via heterochromatin formation,[8–10] in gene activation by facilitating the formation of super-enhancers,[11] in buffering cellular noise,[12] in modulating enzymatic reactions[13] and in sensing pH changes in the skin.[14] However, some biomolecular condensates emerge spontaneously inside cells without as-yet clearly identified functions; it has been hypothesised that some of these might be implicated in the emergence of phase-separation-related pathologies.[15] Indeed, aberrant LLPS of the proteins Fused in Sarcoma (FUS) and Tau has been associated with the onset of degenerative diseases such as amyotrophic lateral sclerosis and Alzheimer's disease, respectively.[15] More recently, biomolecular condensates have been proposed as promising new tools to partition anti-cancer drugs preferentially to cancer cells.[16] Such a richness of behaviours highlights the importance of learning to design protein mutations that can alter the stabilities of condensates.

When designing protein mutations, it is useful to consider that the thermodynamics of phase separation is driven by the competition between interaction enthalpies and the entropic favourability of mixing.[17–19] In the context of protein solutions at physiological conditions, LLPS is principally stabilised by π-stacking and cation–π interactions, followed by charge–charge, dipole–dipole and other hydrophobic interactions.[20–23] The relative contributions of different amino-acid pair interactions to LLPS stability is further modulated by the experimental conditions, including the temperature, pH and salt concentration.[21] In biomolecular systems, the multivalency in mixtures is thus the main physical parameter that defines the ability of a system to undergo LLPS:[6,15,22,24–26] biomolecules with higher valencies can establish a larger number of weak attractive interactions with other species and hence form a more stable condensate.

The connections between LLPS and cellular function, and between aberrant LLPS and human pathologies, suggest that learning how to control or even prevent the phase separation of proteins by subtly mutating their amino-acid sequences would be highly desirable. However, the sequence space of even the smallest naturally occurring proteins is immense, which makes the task of choosing mutations manually extremely inefficient; what is more, even if small-scale modifications of a single protein that promote phase separation might be possible to design manually with some physical intuition, biomolecular LLPS is a

[a)]Present address: Department of Biochemistry, University of Oxford, 3 South Parks Road, Oxford, OX1 3QU, United Kingdom
[b)]E-mail: rc597@cam.ac.uk
[c)]E-mail: ar732@cam.ac.uk

collective phenomenon involving many weak interactions, and it is not at all straightforward to anticipate how small sequence modifications affect the phase behaviour of a protein mixture without the use of an algorithm.

Indeed, optimising biological LLPS is especially difficult because *in vivo* biomolecular condensates can be highly multicomponent systems.[25,27–29] Furthermore, over 270 distinct multivalent proteins have been shown to undergo LLPS *in vitro*.[30] Despite this complexity, the properties of condensates can be successfully approximated in vitro by considering just a fraction of biomolecules, known as 'scaffolds', since such molecules tend to dominate the phase behaviour:[6,24,31] the addition of biomolecules that are recruited to condensates via their interactions with scaffolds, termed 'clients',[6,24] impacts the stability of condensates only marginally,[26] making the problem somewhat more manageable.

A wide body of work has significantly advanced our understanding of how changes in amino-acid sequence transform the phase behaviour of different proteins.[22,23,32–38] Notably, tightly integrated experiments and simulations with the 'stickers and spacers' model explain how changing the number and patterning of aromatic and charged residues can alter the phase diagrams of prion-like-domain proteins in a predictable manner.[22,23] Experiments also demonstrate that multimerising the arginine/glycine-rich RGG domain of LAF1 leads to controllable phase separation,[39] and minimal coarse-grained simulations of point mutations of two designer proteins exemplify how on-demand modulation of protein phase behaviour can be achieved *in vivo*.[40]

Alongside globular domains, intrinsically disordered regions (IDRs) are thought to be one of the main drivers of LLPS in protein systems.[41,42] Various theoretical approaches, including random-phase approximation theory, have been applied to LLPS of IDRs.[43,44] Such treatments can rationalise aspects of charge distribution in phase-separating IDRs, and their relative simplicity has allowed us to gain significant intuition for the electrostatic aspects of phase separation. However, the scope of theory is limited by the inherent approximations that are of necessity made and the sheer complexity of biological multicomponent systems. More realistic representations of proteins are usually too complex for analytical treatment, but can be studied in computer simulation, giving us molecular-level insight into the phase behaviour of biological systems. Computational approaches hold significant promise of enabling us to probe amino-acid sequence space efficiently and to design protein mutations which enhance or inhibit LLPS of a protein solution. Although all-atom simulations of LLPS in explicit solvents are only now slowly becoming computationally tractable[45–48] due to the exponential scaling of search space with system size, the ABSINTH ('self-Assembly of Biomolecules Studied by an Implicit, Novel, Tunable Hamiltonian') all-atom and implicit solvent framework[49,50] of Pappu and colleagues has been consistently applied to estimate relative critical solution temperatures from single-molecule properties and to capture subtle variations in sequence space in agreement with experiment.[22,51] Most recently, ABSINTH has been successfully paired with Gaussian cluster theory to compute full sequence-dependent phase diagrams of proteins.[51]

Theory and simulations using coarse-grained potentials – from patchy particles to lattice models[22,26,35,39,45,49,52–58] – have provided significant microscopic insight into the physics of phase separation. Amongst the growing body of coarse-grained models available to investigate protein phase behaviour computationally, a noteworthy approach is the 'stickers-and-spacers' model of Pappu and colleagues. The stickers-and-spacers model represents multivalent proteins as heteropolymers composed of stickers (LLPS-binding motifs) and spacers (regions in between stickers), and has been used to generate phase diagrams of proteins in perfect agreement with experiment and to elucidate the underlying molecular driving forces.[22,36,55] The residue resolution HPS ('HydroPhobicity Scale') and KH ('Kim–Hummer') coarse-grained models of the Mittal and Best groups,[35,59] combined with the direct-coexistence simulation method, also stand out among the techniques to assess the effect of amino-acid sequence variation on protein LLPS. Some of the advantages of the HPS/KH models include their transferability and their ability to compute phase diagrams of large proteins (i.e. up to ~500 residues per protein) at residue resolution.

Here, we develop a genetic-algorithm approach coupled to the sequence-dependent coarse-grained model of proteins with amino-acid resolution of the Mittal and Best groups[35] [Fig. 1(A)(i)] to design protein mutations that can enhance or inhibit LLPS, and use the all-atom implicit solvent ABSINTH framework[49,50] to verify the validity of our predictions. Our method takes advantage of the computational efficiency of the residue-resolution coarse-grained models to search sequence space, and the higher accuracy of ABSINTH to predict experimentally-consistent critical temperature of proteins. Genetic algorithms have been used since the early 1990s with considerable success in a variety of fields, from reaction dynamics[60,61] to crystal and cluster structure prediction,[62,63] protein evolution[64–66] and drug design,[67] including when coupled with computer simulations.[68–73] In the past year, the integration of genetic algorithms and coarse-grained models applied to biological questions seems to be gaining traction;[73,74] indeed, very recently, a genetic-algorithm approach has been used to design sequences of proteins that exhibit lower critical solution temperatures.[74] In our implementation, we anchor a genetic algorithm to a fitness function that is fast enough to be evolved and that represents a good proxy for the critical solution temperature, which measures the ability of a protein to phase-separate. With this approach, we systematically evolve the amino-acid sequences of the IDRs of three naturally occurring proteins that are known to phase-separate *in vitro* via homotypic interactions, and we show that we can drive the genetic algorithm either to enhance or to inhibit their LLPS. By shuffling the amino-acid sequences in chunks of varying lengths, we also identify the binding domains of the IDRs that are essential to drive LLPS (the 'stickers') and the connecting regions (the 'spacers').[22,36,55] By investigating LLPS in the vicinity of known phase-separating sequences, we can infer which features of a sequence drive phase separation in biological systems. While previously, artificial sequences have been probed in a systematic way, for instance in the context of charge patterning,[45] our work also complements very recent results obtained on LAF1-IDR[39] and Ddx4-IDR.[58] Although some of the fine features of our findings may be model-specific, we validate the robustness of our genetic algorithm by repeating core runs using an alternative parameterisation of the protein coarse-grained model.[58] We also benchmark the model's capacity to predict critical solution temperatures against experimental data[23] and the ABSINTH implicit solvation model and all-atom force-field paradigm.[49,50]

II. RESULTS

A. Protein phase behaviour can be guided by a genetic algorithm

A free choice from amongst the 20 canonical amino acids in a protein with n residues amounts to an n-dimensional vector

with 20^n possible sequences, where for each sequence one might attempt to compute some property that characterises the sequence's LLPS behaviour. One possible quantity that can serve this purpose is the upper critical solution temperature T_c, above which no de-mixing occurs. However, an exhaustive search of sequence space would be prohibitively expensive. Moreover, finding the 'optimal' critical temperature, however we might choose to define it, is not itself the aim. For example, a poly-F chain has a particularly high T_c (see SI Sec. S2) in the protein model we have used, but studying it in detail is not particularly helpful in understanding what drives biological phase separation. We instead focus on biologically occurring proteins and evolve their amino-acid sequences with the aim of finding individual examples of sequences that either extend or narrow (as opposed to maximise or minimise) the range of thermodynamic conditions where homotypic LLPS occurs; in other words, our goal is to use the genetic algorithm to perform only local optimisation. In particular, we are interested in the effect of relatively small changes to the amino-acid sequence on phase separation, as these are instrumental to understanding how modifications can be designed to control the phase behaviour of proteins *in vivo*. Moreover, such modified sequences might more easily be introduced into cells.

To determine sets of mutations that shift the phase behaviour of a protein in the desired direction, we start from a reference amino-acid sequence and perform direct-coexistence molecular dynamics (MD) simulations of a sufficiently large number of copies of that protein (see Fig. 1(A)(ii) and SI Sec. S1.2). In direct-coexistence simulations, we explicitly simulate two different phases – a protein-enriched solution and a protein-depleted solution – in contact with each other in the same simulation box. By performing such simulations at several temperatures, we can approximate the compositional phase diagram, which indicates which phase is thermodynamically stable as a function of temperature and protein density. We then seek to evolve the protein towards enhanced or inhibited LLPS by developing a genetic algorithm (see Methods) that iteratively proposes stochastic amino-acid sequence mutations, selecting a few at each iteration amongst those that induce the strongest effect amongst the set in the protein LLPS, and mutating again. For a given protein model, whether such a genetic-algorithm approach can succeed in evolving protein phase behaviour depends on the quality and efficiency of the fitness function used to control it. Our genetic algorithm uses the difference in composition densities of the protein-poor and protein-rich phases at constant volume as its fitness function. As we show below, our fitness function is both computationally inexpensive and a good metric to determine whether a set of mutations would result in enhanced or inhibited LLPS. Although the critical solution temperature of a protein mixture may seem like an obvious order parameter to determine whether a specific set of mutations promotes or inhibits LLPS – by raising and lowering the critical solution temperature, respectively – computing it in every round of the evolution process when using a residue-resolution coarse-grained protein model is computationally infeasible. This is because estimating the critical solution temperature requires an evaluation of a full phase diagram of the protein solution, which in turn requires either the use of very expensive free-energy methods,[75] or performing direct-coexistence simulations at a number of different temperatures, each involving long MD simulations of a large number of copies of the same protein, and analysing the results to extrapolate the data and estimate the critical temperature. By contrast, evaluating the difference in composition densities requires only one set of direct-coexistence simulations to be run at a fixed sub-critical temperature (i.e. below T_c).

B. The case of the PLD of FUS

1. *The range of stability of LLPS can be evolved.*

As an initial model system, we investigate the behaviour of the prion-like domain (PLD) of the FUS protein, an IDR rich in tyrosines and mostly devoid of charged residues. Although the PLD of FUS only phase-separates *in vitro* at somewhat extreme conditions with respect to the physiological ones (namely low salt concentrations of 37.5 mM NaCl and high protein concentrations of 6 µM to 33 µM),[76] PLD–PLD interactions and PLD–arginine-rich domain interactions drive LLPS of the full FUS protein under physiological conditions, both *in vitro* and in cells.[77] We first use direct-coexistence molecular dynamics simulations to approximate the compositional phase diagram (i.e. in the temperature versus protein-density space) of the PLD of FUS. We then seek to evolve the system's critical solution temperature by introducing our genetic algorithm (see Methods), which allows us to mimic, broadly speaking, the evolutionary pathways that might drive phase separation in nature. Starting from the reference amino-acid sequence of FUS PLD ('WT-FUS') given in SI Sec. S8, we use our genetic algorithm to attempt separately to increase and to reduce the width of the binodal curve of the compositional phase diagram. We show the population fitness [Eq. (2)] and the diversity of the population as functions of the genetic-algorithm round in Fig. 1(B) for the case of increasing the phase diagram width; analogous results for the case of reducing it are shown in SI Sec. S4. The genetic algorithm is effective in increasing the population fitness in each case, and in both cases the population diversity remains high, indicating that no premature convergence occurs. We also confirm the effectiveness of the driven evolution in our genetic algorithm by contrasting the results to a dummy variant with all the mutagenesis steps intact, but the selection pressure abolished [SI Sec. S3]. Furthermore, over the three repeats of the genetic-algorithm runs we performed to increase the phase diagram width, mutation of all but two residues (27 and 155) was attempted by the genetic algorithm at least once, which indicates a sufficient mutation rate to probe the entire sequence over 20 rounds. In Fig. 1(C), we show that the phase diagram of the evolved FUS PLD in the former case exhibits a large increase in the range of temperatures and densities at which LLPS occurs, with the critical temperature increasing by ~65 % compared to the WT-FUS PLD. Although we have used the width of the phase diagram as a proxy for the critical solution temperature, Fig. 1(C) confirms not only that the critical solution temperature behaves in the expected way, but also that genetic algorithms with simple fitness functions can significantly perturb the LLPS behaviour, leading to an effective gradient in sequence space. These results suggest that a genetic algorithm can be used to search the sequence space of proteins efficiently and can help propose sequence mutations that yield meaningful changes in the proteins' compositional phase diagrams. Importantly, as mentioned earlier, our approach is computationally tractable because our goal is to find candidate mutations that can purposely increase or narrow the range of stability of LLPS, rather than to identify the specific amino-acid sequences that give rise to a true maximum or minimum critical solution temperatures, and hence, performing just a few iterations of the algorithm is sufficient.

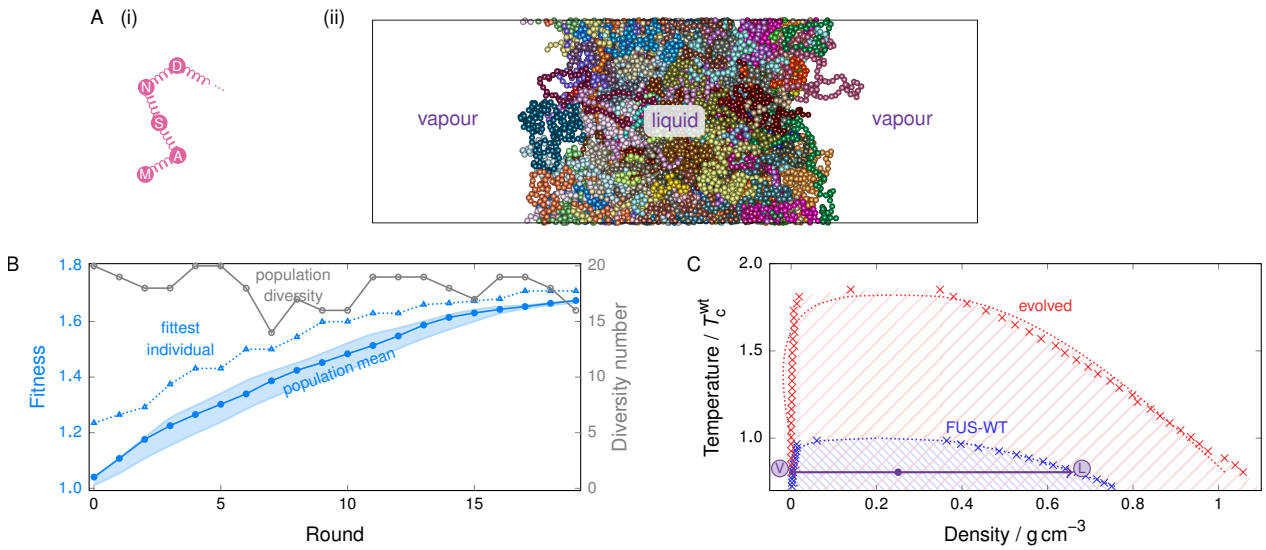

FIG. 1. **Evolution behaviour of FUS.** (A) (i) A schematic representation of the model used. Each amino acid is represented by a bead, and beads are connected with harmonic springs. (ii) A snapshot of a typical simulation cell exhibiting coexistence between a liquid-like (protein-rich, high-density) fluid and a vapour-like (protein-poor, low-density) fluid. The box is periodic in all directions. Different colours are used to represent beads in different protein chains. (B) Typical genetic-algorithm progression for FUS where the fitness function *increases* the width of the phase diagram. The fitness function [Eq. (2)] increases by ~65 % over 20 rounds. The fittest individual is 5 % to 20 % fitter than the mean in most rounds. The population diversity, i.e. the number of distinct sequences present in the overall population of 20, remains high throughout the run. The shaded area corresponds to the range of values of the mean fitness obtained from 3 independent genetic-algorithm runs. (C) Comparison of representative phase diagrams before and after genetic-algorithm runs, confirming that the fitness function choice was suitable. Pale hatched lines indicate the approximate region of phase separation for each case. Error bars in the density evaluations are smaller than the symbols, and dotted lines are fits as detailed in SI Sec. S1.2. The point labelled in violet corresponds to the snapshot shown in panel (A), with the densities of the vapour-like and liquid-like fluids labelled 'V' and 'L', respectively. All densities reported are for the protein in a massless implicit solvent.

2. LLPS evolution can be driven by changes in hydrophobic and aromatic residue composition.

We analyse the extent to which all interactions other than direct charge–charge interactions, which here we term collectively as 'hydrophobicity', govern the evolution of the phase behaviour of proteins by estimating each amino acid's relative degree of hydrophobicity. We use the hydrophobicity scale proposed in Ref. 78, which is quantified as the λ parameter in the coarse-grained model of Dignon *et al.*,[35] and which can be used to scale the well depth of the modified Lennard-Jones potential in an amino-acid-specific way (see SI Sec. S1.1). In Fig. 2(A), we show the amino-acid compositions of the populations resulting from genetic-algorithm runs in which T_c is increased and runs in which it is decreased, broken down by amino acid and ordered by the extent of hydrophobicity, alongside the reference WT sequence. The amino-acid sequences of the WT of the FUS PLD and examples of its evolved analogues are given in SI Sec. S9. In the case of runs targeting an increase in T_c, there is a general shift towards higher hydrophobicity, whilst the case where T_c is targeted to decrease shows a trend towards highly polar and charged amino acids. These trends in amino-acid composition confirm that, even though there are more strongly hydrophobic than weakly hydrophobic amino acids available for insertion, evolution of the FUS PLD is able to be driven in both the hydrophobic and hydrophilic directions. [One specific limitation of using a coarse-grained potential on results from a genetic-algorithm framework is its broadening effect on amino-acid composition: physically more important amino acids can stochastically be replaced by less important ones simply because their force-field parameters are similar. As a result, when we describe the strength of 'hydrophobic' interactions, we refer to the interactions in general between non-charged amino acids. However, a more accurate model to describe residue–residue interactions could in the future be coupled to our genetic algorithm to resolve such interactions in more detail.]

In addition to the attractiveness of the hydrophobic interactions, a further factor determining the strength of hydrophobic interactions is the size of each amino acid. We quantify this by σ, the Van der Waals radius of each amino acid (see SI Table I). In Fig. 2(D), we show that the average size of the amino acids in the sequence population increases as a function of the genetic-algorithm round, implying that the average size of the amino acids increases through evolution. Furthermore, as shown in Fig. 2(B)(iii), although the effects on hydrophobic attractiveness and σ values largely correlate at most residues of the protein sequence (with a Pearson coefficient of 0.42), this is not invariably the case, giving us the first indication that the size of the amino acids could play an independent role in determining LLPS properties. Although the overall increase in amino-acid size might be explained at least in part by the amino-acid size defining the range of the hydrophobic attractions, we hypothesise that the main physical driving force explaining the increase of both hydrophobicity and size is the ability of larger and more hydrophobic amino acids to form a more densely connected, and in turn more stable, condensed liquid-like protein-rich phase.[26] To test this hypothesis, we compute the pair correlation function of the protein-rich phase for both FUS-WT and one of the evolved sequences at a common number density [Fig. 2(C)]. The nearest-neighbour maximum is 13 % higher than in the wild type, indicating a greater degree of local structure and an increase in the number of nearest-neighbour beads compared to the WT, which has previously been shown[26] to correspond to a greater protein valency.

Because the sequence of the WT PLD of FUS only contains two (negatively) charged amino acids, its LLPS must be driven by hydrophobicity. However, our results show that the FUS sequence lies on a hydrophobic gradient in sequence space:

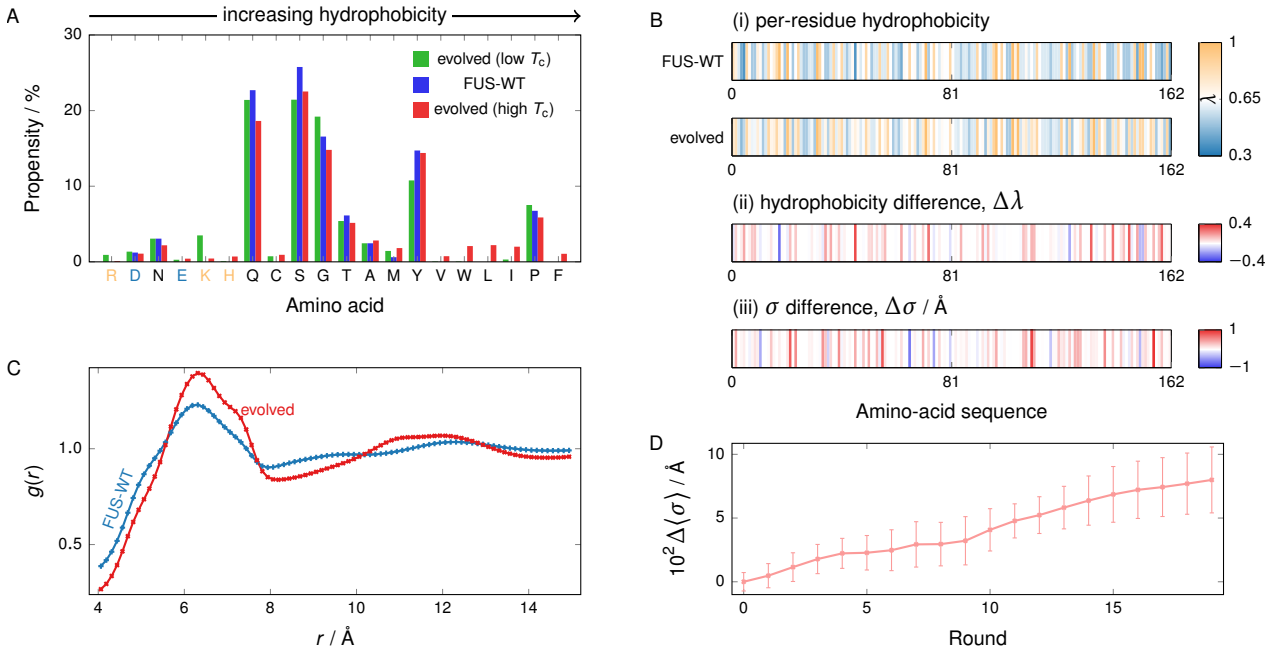

FIG. 2. **Properties of evolved FUS.** (A) Amino-acid composition before and after applying the genetic algorithm to increase and reduce the phase diagram width, and hence the critical temperature, starting from WT-FUS. Amino acids are plotted in order of increasing hydrophobicity [see SI Table I]. Positively charged amino acids are indicated in light orange and negatively ones in light blue. In genetic-algorithm runs which increase T_c, hydrophobic amino acids become favoured, whilst the converse holds for genetic-algorithm runs which decrease T_c. (B) Map of (i) the per-residue hydrophobicity along the sequence, (ii) the change from the wild type to the evolved protein after evolution towards higher T_c and (iii) the change of the per-residue amino-acid size (σ). The data for the evolved protein are averaged over the entire population at the end of 3 independent genetic-algorithm runs. No larger-scale regional preference for modification is readily apparent. Trends in hydrophobicity and σ are largely correlated. (C) Comparison of the pair correlation function $g(r)$ before and after evolving FUS with a genetic algorithm towards a larger phase diagram width. The data were computed at the same bead number density $N/V = 7.0\,\mathrm{nm}^{-3}$ and temperature $T = 0.8T_c^{\mathrm{wt}}$. In both cases, this temperature is below the critical point and the density corresponds to the protein-rich (liquid-like) phase, i.e. a point that lies above the binodal line on the phase diagram. We compute $g(r)$ by finding for each bead i in the system the number of all non-harmonically bonded other beads within a distance $r + \delta r$ of the each other for bins of width δr, averaging over each bead i in the system, and normalising the result by the volume element and the (common) number density. The symbol size is larger than the standard deviation of the average across 4 independent simulations. In the case of the evolved system, the more pronounced nearest-neighbour maximum indicates the local environment is more structured than in the case of WT-FUS. (D) The average σ value of the amino acids in the population increases over the course of the genetic-algorithm run. Error bars are standard deviations of the averaged σ value of individual sequences with respect to the pooled population from the three genetic-algorithm runs.

an increase in hydrophobicity effects an increase in the critical solution temperature. It has been proposed[79,80] that the driving force for the LLPS of this FUS IDR is specifically the interactions between tyrosine residues dispersed through the sequence. Although such interactions are only implicitly captured in the coarse-grained protein model we use through its hydrophobicity parameter, and thus the distribution of amino acids obtained follows broad trends rather than converging to a distinct amino acid or motif, our results are consistent with previous work,[79,81,82] and meaningful trends in composition and sequence can be observed from our simulations. Our results thus suggest that evolving a protein sequence which is dominated by hydrophobic residues, as is the case for the PLD of FUS, towards enhancing its propensity for LLPS is efficiently achieved by protein mutations that increase the average attractiveness and size of the protein's uncharged residues.

3. *Hydrophobic patterning has a minimum length scale.*

In Fig. 2(B), we show how the accumulated changes in sequence, represented as the hydrophobicity of a residue, map onto the WT-FUS sequence. There are no larger-scale regions along the sequence where modifications occur preferentially; instead, there appears to be a stochastic increase in hydrophobicity, with less hydrophobic residues being replaced by more hydrophobic ones. However, since short runs of the genetic algorithm cannot result in perfectly uniform replacement attempt probabilities, we cannot expect to be able to resolve small-scale features in amino-acid sequence space. In order to investigate such small-scale features, we therefore first 'shuffle' the sequence without changing its overall amino-acid make-up. We choose chunks of varying lengths by randomly choosing two positions along the sequence and exchanging chunks of l amino acids starting from those two positions. The ends of the sequence are treated periodically to ensure no positional selection bias against the ends. The swapped amino acids may be of the same type. We record the fitness function of the protein sequences as a function of the total number of amino-acid pairs changed, i.e. the number of exchange steps multiplied by l. In Fig. 3(A), we show the variation of the fitness with chunk length after ~100 amino acids have been shuffled. The error bar, which shows the standard deviation across several independent runs, is a useful measure of the sensitivity of the fitness function to amino-acid sequence. Very small chunk lengths, particularly of 1 or 2 amino acids, are highly disruptive to phase separation, while larger chunk lengths only cause smaller modulations. From these results, we can conclude that segments of 2–3 successive amino acids are crucial in driving LLPS in the PLD of FUS, representing the length scale of some sequence feature. To investigate the nature of this feature, we repeated the shuffling analysis with a hydrophobicity bias, where only the most

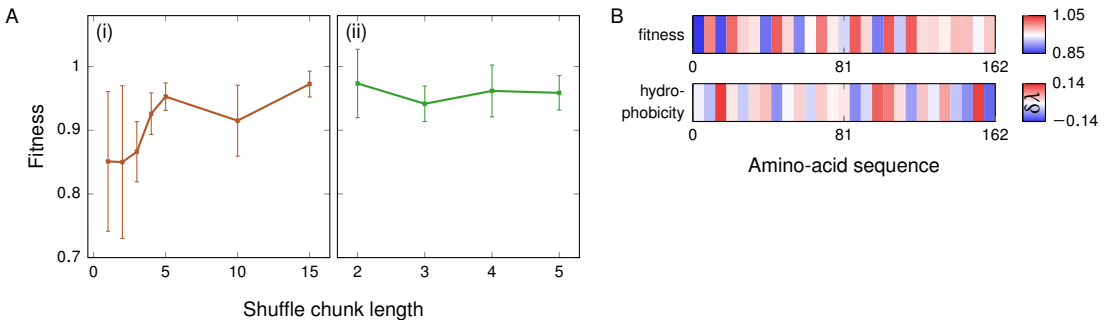

FIG. 3. **Determining the characteristic length scale.** (A) (i) The fitness of the system as a function of the chunk length following the shuffling of approximately 100 amino acids. Error bars show the standard deviation across several shuffling runs. There is a significant difference in the fitness and the error bars for small chunks of 1 and 2 amino acids, whereas the error bars for the larger chunk sizes are considerably smaller. (ii) Analogous results for shuffling runs with a hydrophobicity bias, where exchanges were allowed only amongst the top 30 % of chunks by hydrophobicity. Representative fitness functions as a function of the number of amino-acid pairs shuffled are shown in SI Fig. S5. (B) The value of the fitness function (relative to the WT fitness) when chunks of 6 amino acids of FUS-WT are separately replaced with glycine, and the average hydrophobicity of these same 6-residue chunks of FUS-WT relative to the hydrophobicity parameter of glycine, $\delta\lambda = \langle\lambda\rangle_{\text{chunk}} - \lambda(\text{G})$, where $\lambda(\text{G}) = 0.649$. Where glycine represents a gain in hydrophobicity, the fitness change is largely positive, and vice versa.

hydrophobic of all possible contiguous chunks are exchanged. The dependence on chunk size largely disappears, implying that it was small hydrophobic patches that were previously disrupted by shuffling [Fig. 3(A)(ii)]. The phase separation therefore appears to be governed by a hydrophobic patterning of a minimum length scale of 2–3 amino acids. This is consistent with the 'stickers-and-spacers' paradigm of phase-separating proteins, in which proteins are considered to comprise stickers – corresponding to the attractive protein regions that drive LLPS, in our case the small chunks of 2–3 amino-acid residues – that are connected by less attractive regions termed spacers.[22,36,55]

Once we know this minimum length scale, we can investigate the effect of hydrophobicity by replacing successive chunks of the amino-acid sequence with a fixed amino acid. The conventional approach to probe the function of specific residues is alanine scanning.[83–85] As we are interested in how hydrophobicity affects phase behaviour, in this case, we mutate amino acids to glycine rather than alanine, as the former has the median hydrophobicity in the coarse-grained protein model (see SI Table I). Although in experiments or all-atom simulations, such a replacement may be less appropriate, as glycine disrupts protein secondary structure by its dihedral angle preferences,[86] in the CG model this effect is immaterial, as no conformational terms are considered. We replace successive chunks of 6 amino acids in each case; this chunk size is about 2–3 times the characteristic length scale of hydrophobic patterning, ensuring that differences observed most likely arise from the overall difference in hydrophobicity rather than a disruption of localised 'stickers'. Fig. 3(B) shows the results of a glycine scan projected onto the chunk-averaged hydrophobicities of the WT protein. The curves anti-correlate for most of the sequence (with a Pearson coefficient of −0.5), reflecting that in hydrophobic stretches, mutating to glycine decreases hydrophobicity and thus decreases fitness, while the converse holds for hydrophilic stretches, confirming the dominance of hydrophobicity as a driving force for LLPS in this case.

C. Charge patterning may be an alternative driving force for evolution of protein phase behaviour

Not all proteins that exhibit LLPS are expected to be governed by the same driving force. For example, the patterning of charges has been suggested to contribute to LLPS in charge-rich proteins,[44,45] while the phase separation of the intrinsically disordered region of the protein hnRNPA1, which belongs to the family of heterogeneous nuclear ribonucleoproteins, has been shown to be driven by the interaction between linearly dispersed aromatic residues within the polar sequence.[22,23,34]

Here, we will use the hnRNPA1 IDR as a test case to complement the behaviour observed for FUS. We define the IDR of hnRNPA1 as its first 135 residues, 11.9 % of which carry a formal charge, and which has been shown to phase-separate *in vitro*.[76] We used a genetic-algorithm-driven evolution of the same fitness function as in the case of FUS; the genetic-algorithm approach results in an increase of fitness whilst maintaining the population diversity [Fig. 4(A)], and the increased fitness function is again a successful proxy for the upper critical solution temperature [Fig. 4(B)]. We also note that in this case, over the three repeats of the genetic-algorithm runs, mutation of all residues was attempted at least once by the genetic algorithm, giving us good coverage of the entire sequence.

The change in amino-acid composition upon applying the genetic algorithm to hnRNPA1 is qualitatively different from the hydrophobicity-driven case of FUS, as hydrophilic/charged residues are not lost and hydrophobic residues appear, statistically replacing some of the highly abundant amino acids of intermediate hydrophobicity [Fig. 4(C)], indicating that the driving force for phase separation may be different from the case of FUS-PLD.

To investigate this further, we have analysed the initial and final populations in the genetic-algorithm runs. As in the case of the PLD of FUS, the hydrophobic attractiveness and the average size of amino acids of the IDR of hnRNPA1 increase to raise the propensity for LLPS [Fig. 4(D)(i–iii)], and as in the case of FUS, the per-residue effects of hydrophobic attractiveness and amino-acid size are largely, but not fully, correlated.

1. *Charge patterning and hydrophobicity can co-evolve . . .*

Additionally, there is a substantial difference in terms of charge patterning over the course of the genetic-algorithm run [Fig. 4(D)(iv),(v)]. Charges are both created and lost across the sequence, but not uniformly so. We show in Fig. 4(E)(i) two measures of the charge patterning, the net charge of a protein

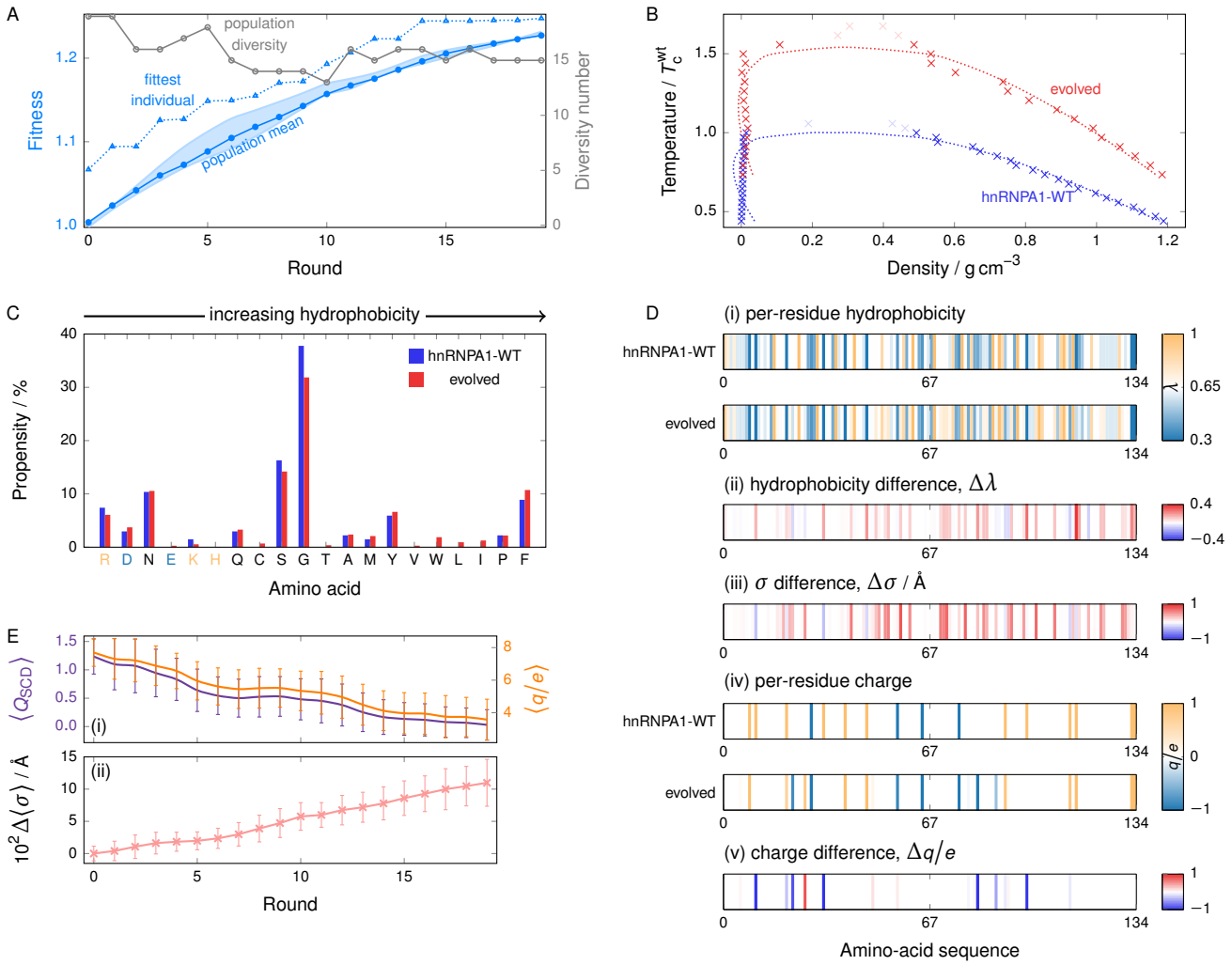

FIG. 4. **Evolution of hnRNPA1.** (A) Typical GA progression for hnRNPA1. The fitness function [Eq. (2)] was evaluated at $T = 0.57\,T_c^{wt}$, and increases by ~20 % over 20 rounds, while maintaining population diversity. The shaded area corresponds to the range of values of the mean fitness obtained from 3 independent genetic-algorithm runs. (B) Comparison of representative phase diagrams before and after genetic-algorithm runs, showing a ~50 % increase in critical temperature. Dotted lines are fits, and greyed-out points lie above the critical point, as detailed in Section SI Sec. S1.2. (C) Amino-acid composition before and after using a genetic algorithm to increase the phase diagram width for hnRNPA1, revealing the appearance of hydrophobic amino acids, while no bias against charged or polar amino acids is observed. Amino acids are plotted in order of increasing hydrophobicity [see SI Table I]. Positively charged amino acids are indicated in light orange and negatively ones in light blue. (D) Map of (i) the per-residue hydrophobicity along the sequence, (ii) the change from WT hnRNPA1 to the evolved protein, (iii) the per-residue change σ value and (iv), (v) analogous maps for the charge. The data for the evolved protein are averaged over the entire population at the end of 3 independent genetic-algorithm runs. Partial charges reflect only partial carriage in the population. Some charges have appeared and some have disappeared; the overall balance is towards charge neutralisation. (E) (i) The sequence charge decoration (SCD) and the charge number (q/e) of the population decrease over the course of the genetic-algorithm run, indicating an evolutionary charge neutralisation. (ii) The average σ value of the amino acids in the population increases over the course of the genetic-algorithm run. Error bars are standard deviations of the averaged σ value of individual sequences with respect to the population.

chain, and the sequence charge decoration (SCD), defined as[87]

$$Q_{\mathrm{SCD}} = \frac{1}{N} \sum_{m=2}^{N} \sum_{n=1}^{m-1} q_m q_n (m-n)^{1/2}, \qquad (1)$$

where q_i is the formal charge number of residue i and N is the length of the amino-acid sequence. Q_{SCD} has been shown to anti-correlate with the upper critical solution temperature of an IDR.[45] These two parameters show a virtually identical evolution through the genetic-algorithm run, which indicates that, in this case, the decrease in charge separation as measured by Q_{SCD} results from a net decrease in the overall charge of the protein. Specifically, this arises from the creation of a larger number of negative than positive charges.

A considerable amount of work has already been done in the context of the role of charge patterning.[39,44,88–91] Although it is perhaps not overly surprising that a more even distribution of positive and negative charges allows for the largest number of attractive interactions, which in turn drives the formation of liquid-like phases, we show below that the precise nature in which sequences evolve depends on the medium in which the proteins of interest exist.

The local gradient in sequence space around hnRNPA1-WT has components in both hydrophobicity and charge redistribution. However, in the literature, hnRNPA1 LLPS is not commonly associated with charges.[22,34] This prompts the question of how important the factors are in absolute terms. A crude estimate can be obtained from an analysis of the components of the pairwise energy in our simulations, shown in Table I. In particular, we split up the energy into a 'hydrophobic' (LJ) contribution and a Coulomb (electrostatic) contribution. Both components become more favourable over the course of a genetic-algorithm run, indicating that the sequence-space gradi-

ent towards higher T_c encompasses both charge rearrangement and hydrophobicity. These effects operate in parallel, but the hydrophobicity component contributes considerably more to the attractive energy in absolute terms.

2. … but need not necessarily, even in charge-rich sequences.

To check the applicability of the two mechanisms driving the evolution of the capacity to undergo LLPS that we have identified to other protein sequences, we have also investigated an IDR of the protein LAF1, which is a DDX3-family RNA-helicase found enriched in *C. elegans* P-granules, in which it drives phase separation. The IDR we have focussed on has been shown to be both necessary and sufficient for LLPS.[92] It contains a significant proportion of charged amino acids, with 22.4 % of its 168 residues carrying a formal charge. This IDR has also been shown to phase-separate in simulations of the CG model used here,[35] and a recent study[39] combining CG simulations, all-atom simulations and turbidity assays has identified a sticky hydrophobic stretch as well as tyrosine and arginine residues to be involved in LAF1 LLPS. Additionally, it has been suggested[39] that the even distribution of charges across the sequence may suggest that charge patterning is a controlling determinant of LLPS. Simulations and *in vivo* experiments have been carried out in corroboration of this hypothesis.[39]

In order to compare the behaviour of LAF1 to the two cases already considered, we have evolved its sequence using the same genetic algorithm. As before, we have computed the phase diagram at the end of the genetic-algorithm run [Fig. 5(B)]. Although the simulations are slower with this system and finite-size effects are more pronounced, the genetic algorithm with this fitness function can successfully increase the critical solution temperature [Fig. 5(A–B)]. Since the higher computational cost restricted us to only one shorter run of the genetic algorithm, there was incomplete coverage of the sequence with mutations, in contrast to FUS and hnRNPA1. While this is not an issue for interpreting our results, since complete coverage is not essential to observe the trends we look for, we highlight those residues which were not touched by the genetic algorithm for clarity [Fig. 5(D)]. Compared to both FUS and hnRNPA1, the composition of the resulting evolved sequence population is less significantly changed [Fig. 5(C)], although this is consistent with the fact that the fitness function increases more slowly and the overall critical solution temperature is only ∼30 % higher than the wild type in the simulations considered (compared to 65 % and 50 % for FUS and hnRNPA1, respectively). Nevertheless, there is a limited increase in hydrophobicity [Fig. 5(D–E)], with no region particularly favoured in terms of increased hydrophobicity, even though the amino acids early in the sequence (i.e. nearer the N-terminus) are on average more hydrophobic than later ones. The change in the range of hydrophobic interactions, however, as quantified by the average σ values [Fig. 5(E)], is more significant ($\Delta\sigma$ = 0.0396 Å over 10 rounds). This is comparable to the level of change we observed in FUS ($\Delta\sigma$ = 0.0322 Å after the first 10 rounds). In particular, almost all changes made to the sequence by the genetic algorithm lead to larger amino-acid sizes, even though in terms of hydrophobic attractiveness their effects are much more varied. This leads us to speculate that the extended range of attractive interactions may be the dominant factor in driving the evolution of hydrophobic interactions in this case, rather than the λ values, which change less.

The change in charge [Fig. 5(D)(v)] is also relatively modest, and mainly entails the loss of existing net charge [Fig. 5(E)(i)]. This is consistent with charge segregation, as quantified by the sequence charge decoration parameter, also shown in Fig. 5(E)(i), which similarly decreases over the course of the genetic-algorithm run, but whose decrease is significantly less pronounced than in the case of hnRNPA1.

TABLE I. Changes in contributions to the average pairwise energy per bead between the WT and an evolved sequence of FUS, hnRNPA1 and LAF1. Standard deviations for the simulation averages are given in brackets and apply to the least significant figure. For LAF1, the evolved population is diverse in terms of these changes, and two representative examples are shown, labelled *[a]* (fitness 1.34) and *[b]* (fitness 1.28). For FUS , sequence evolution results in a change almost exclusively to the hydrophobic part of the pairwise energy. For hnRNPA1 and LAF1*[a]*, both Coulomb and hydrophobic interactions are more favourable in the evolved sequences, but hydrophobic interactions contribute more in absolute terms. For LAF1*[b]*, the Coulomb energy is less favourable in the evolved sequence. All data presented here are obtained at a simulation temperature of 200 K, corresponding to 0.8 T_c(FUS). While the overall average energies themselves depend significantly on temperature, the differences between WT and evolved sequence energies are largely independent of temperature in the range of interest.

System	$10^3\Delta E_{coul}$ / kcal mol^{-1}	$10\Delta E_{LJ}$ / kcal mol^{-1}
FUS	2.2(2)	−4.69(1)
hnRNPA1	−34(1)	−6.1(1)
LAF1*[a]*	−11.5(5)	−2.11(3)
LAF1*[b]*	3.4(4)	−1.04(3)

Similarly to the case of hnRNPA1, we show in Table I the change in the components of the average interaction energy for LAF1 before and after running the genetic algorithm. However, in the case of LAF1, within the final evolved population, different sequences score rather differently in this analysis. Values for two representative evolved variants, termed LAF1*[a]* and LAF1*[b]*, are shown in the table. We have chosen these specific variants as examples of sequences with similar (high) fitness values, but very different contributions to the energy. In particular, LAF1*[a]* behaves similarly to hnRNPA1, with both hydrophobic and Coulomb interactions interactions more favourable in the evolved sequence than in the wild-type, whilst in LAF1*[b]*, the Coulomb energy actually becomes less favourable. Since the genetic algorithm produces sequences in which the Coulomb attractions are enhanced as well as sequences in which they are weakened, charge patterning does not appear to act as an evolutionary driving force in LAF1. Based on these data we conclude that the evolution of the phase behaviour of the LAF1-IDR is primarily driven by hydrophobicity, and in particular by the size of the amino acids in the sequence. Moreover, this result illustrates a useful advantage of the use of genetic algorithms: when phase behaviour can be enhanced in different ways, this can readily be observed, illustrating that the potential energy surface in sequence space in cases such as this has a number of seemingly degenerate or nearly degenerate states.

D. Sequence evolution depends on the composition of the medium

Inside cells, proteins are never isolated, and LLPS in multi-component systems can be significantly different from that in single-component ones.[26,93–95] It is therefore instructive to examine how genetic-algorithm driving behaviour changes upon the addition of a second component to the medium. To this end, we replace one eighth of the chains in the system with a poly-U RNA chain of the same length as the respective protein of interest. Since cation–π interactions are known to be important for RNA–protein interactions,[96] we employ the cation–π model,[58] a reparameterised version of the coarse-grained model we have used so far, alongside RNA–protein

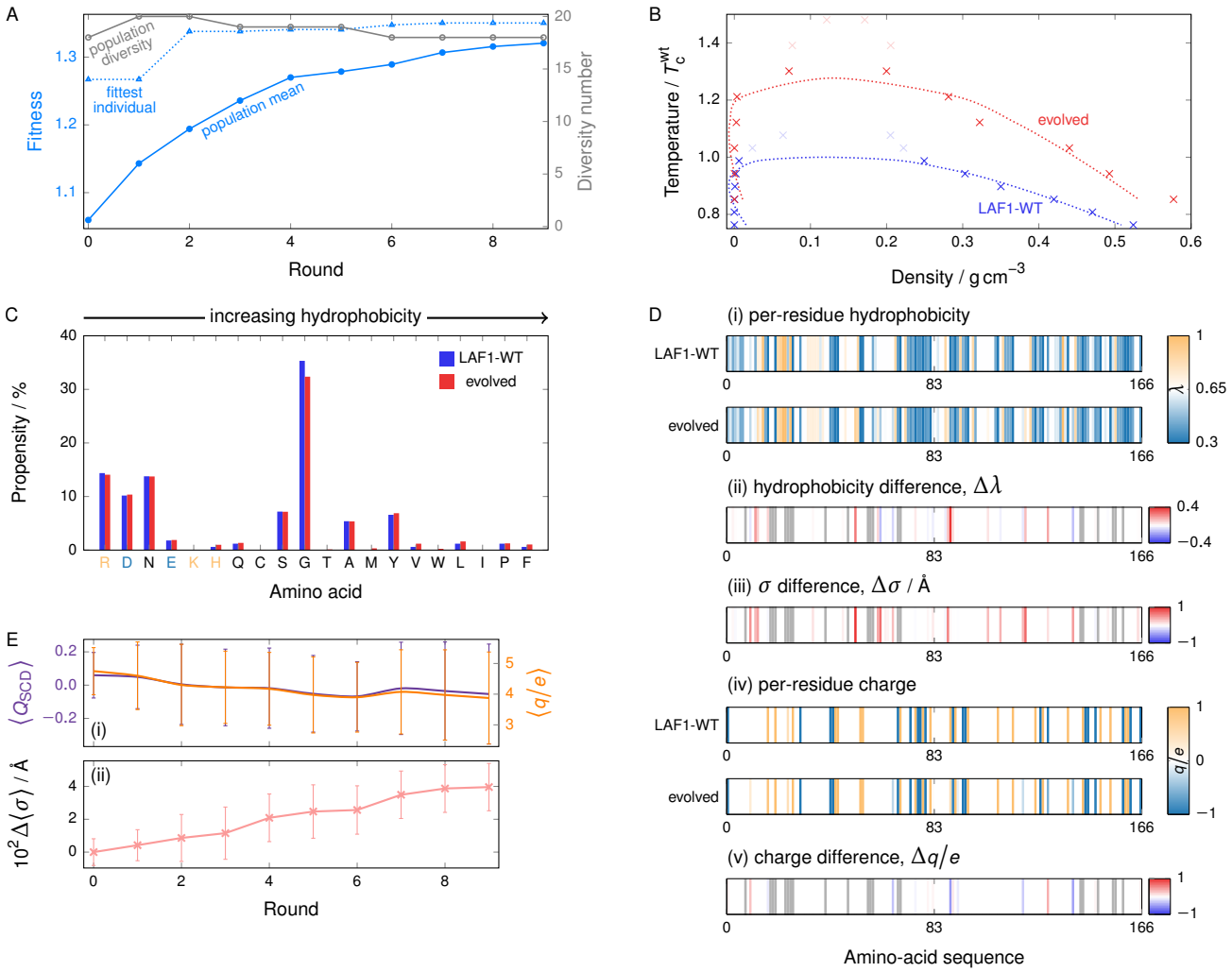

FIG. 5. **Evolution of LAF1.** (A) Typical GA progression for LAF1. The fitness function [Eq. (2)] was evaluated at $T = 0.85\,T_c^{wt}$, and increases by ~30 % over 9 rounds, while maintaining population diversity. To account for thick interfaces, the simulation box was doubled in all directions compared to simulations of FUS and hnRNPA1. (B) Comparison of representative phase diagrams before and after genetic-algorithm runs. Although the phase diagram close to the critical point is especially difficult to equilibrate because of interfacial effects in this system, data points at lower temperatures suggest that the critical temperature increases by ~30 % by the end of the GA optimisation. (C) Amino-acid composition before and after genetic-algorithm runs targeting an increase in T_c for LAF1. There is a slight general increase in hydrophobicity, whilst hydrophilic and charged residues are largely conserved. Amino acids are plotted in order of increasing hydrophobicity [see SI Table I]. Positively charged amino acids are indicated in light orange and negatively ones in light blue. (D) Map of (i) the per-residue hydrophobicity along the sequence, (ii) the change from WT LAF1 to the evolved protein, (iii) the per-residue change σ value and (iv), (v) analogous maps for the charge. Data are shown for one genetic-algorithm run, and those residues where no change was attempted by the genetic algorithm are shown in light grey. Partial charges reflect only partial carriage in the population. There is a slight overall increase in hydrophobicity across the sequence, and there are more charges lost than created during the course of the genetic-algorithm runs. As opposed to the hydrophobicity, the σ values increase for almost all those residues that were changed by the genetic algorithm. More charges are lost than created during the course of the genetic-algorithm runs. (E) (i) The sequence charge decoration (SCD) and the charge number (q/e) of the population decrease slightly over the course of the genetic-algorithm run. (ii) The average σ value of the amino acids in the population increases over the course of the genetic-algorithm run. Error bars are standard deviations of the averaged σ value of individual sequences with respect to the population.

interaction parameters from Regy and co-workers.[59] We discuss this model further in SI Sec. S6. The goal of these tests is simply to determine if our genetic algorithm is sensitive enough to evolve the amino-acid sequence of a given protein so as intentionally to modulate its phase-separating behaviour while taking into account the condensate composition. A poly-U RNA molecule has a high negative charge density and significant scope for non-bonded interactions, which gives it biophysical properties significantly different from the protein it is replacing, making this a good initial test case for more complex multi-component systems.

We show in Fig. 6(A) the change in amino-acid composition at the end of the genetic-algorithm run in systems with and without the additional RNA component, contrasted for (i) FUS and (ii) hnRNPA1. For both proteins, the amino-acid composition change in the presence of a RNA component is different from the single-component case. While the precise changes incurred are likely dependent on the particular model parameters, it is noteworthy that the genetic algorithm can fine-tune composition to enhance phase separation in different ways, depending on the composition of the medium. This applies beyond changes to compensate simply for the introduced charge: in the case of FUS, the overall sequence charge is not drastically affected by the presence of RNA [Fig. 6(B)(i)]; indeed, as can be seen from Fig. 6(A)(i), the creation of more lysine (K) residues is largely offset by creating fewer arginine (R) residues, which have the same charge, and so the net charge increases to a broadly similar extent both with and without RNA present. By contrast, we

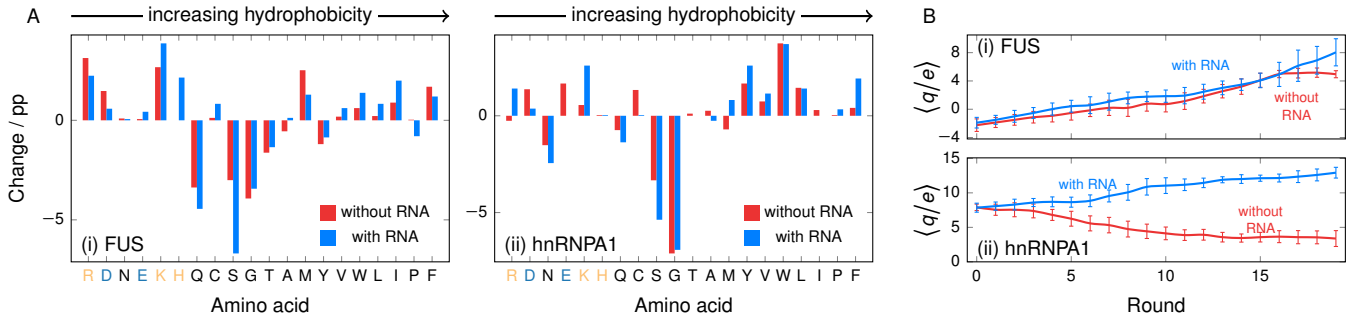

FIG. 6. **Changing the composition of the medium.** (A) Percentage-point difference in amino-acid composition after increasing the phase diagram width for (i) FUS and (ii) hnRNPA1. Amino acids are plotted in order of increasing hydrophobicity [see SI Table I]. Positively charged amino acids are indicated in light orange and negatively charged ones in light blue. The 'with RNA' series corresponds to a system where 1/8 of the chains of the system are replaced with poly-U RNA of the same length as the IDR. (B) Charge content as a function of genetic-algorithm round for (i) FUS and (ii) hnRNPA1. Error bars give standard deviations for the population at each round. The addition of RNA changes the evolution behaviour significantly, particularly in the case of hnRNPA1, favouring higher charge content with the creation of new positive charges, illustrating that the evolutionary driving force depends not only on the initial sequence of the protein, but also on the medium around it.

can observe in Fig. 6(B)(ii) that in the case of hnRNPA1, the addition of RNA affects the evolution in a very different way from that of FUS; namely, the presence of RNA leads to the increase in net positive charge, as opposed to its reduction by the genetic-algorithm run in the absence of RNA. These results illustrate that with the genetic-algorithm approach, we can not only probe the evolutionary driving forces resulting from changing the composition of the medium in which LLPS occurs, but we can also gain insight into how such driving forces depend on the sequence of the protein of interest. In other words, evolutionary driving forces due to the starting sequence and the medium are coupled and evolved together by the genetic algorithm.

Interestingly, all IDRs studied in this work are derived from proteins which in their full-length variants bind to RNA. RNA has widely been studied in the context of LLPS, and can, depending on its concentration, both promote and inhibit protein phase separation,[76] potentially even resulting in re-entrant phase behaviour.[97,98] Therefore, we suggest that extending these preliminary trials on multi-component systems with genetic algorithms could provide insights into the mechanisms by which IDR-containing proteins and RNA might recruit each other.

E. Experimental validation

In order to probe the validity of the trends seen in our genetic-algorithm calculations, and hence the interpretation of the genetic-algorithm predictions, we have computed a series of phase diagrams corresponding to experimental modifications of the IDP of hnRNPA1 as studied in Ref. 23, whose sequences are given in SI Sec. S10. We computed these with two coarse-grained potentials, namely the Kim–Hummer-style parameterisation of the coarse-grained potential of Dignon *et al.*,[35] and the cation–π reparameterisation of the 'HPS' potential introduced by Das *et al.*[58] We show the phase diagrams for these sequences in Fig. 7(A). From the simulation data points, we fit the data to Eqns ((S6)) and (S7), with low-density predictions truncated to zero. From these, we obtain an estimate of the critical temperature in each case. Due to coarse-graining, the absolute temperature scale is of course not directly comparable to experiment; however, if the models are good, one might expect the temperature relative to the critical point to be meaningful. For each sequence, we have therefore estimated the experimental critical temperature from the experimental phase diagrams of Ref. 23, and we show a correlation plot of simulation and experimental data in Fig. 7(B). The linear fits have very reasonable adjusted squared sums of residuals and Pearson correlation coefficients [see Fig. 7], and the coefficient of the linear term has p values of 0.0007 and 0.012, respectively, indicating the statistical significance of the predictor (relative to the null hypothesis of a constant line). The Pearson coefficient for the predictions of the two models against each other is 0.74, and although the majority of data points agree rather well, there are some outliers, particularly for those amino-acid sequences with significant lysine (K) or tyrosine (Y) content. The good positive correlation between the predictions of the sequence-dependent coarse-grained models we have used and the experimental results suggests that we can use these models with success to study broad trends. However, the agreement is by no means perfect for either of these simplified models, demonstrating that there is scope for improving coarse-grained potentials to describe LLPS more accurately.

Next, we probe our final predictions for FUS, hnRNPA1 and LAF1 further by using the more realistic ABSINTH model[49] to estimate the θ temperature of these proteins. The θ temperature of single-molecule coil-to-globule transitions is a well-established proxy for the critical solution temperature of IDR solutions.[38,74,99] It is defined as the temperature at which, for a single protein, there is a coiling transition as evidenced by a sudden change in its radius of gyration as a function of temperature.[51] Linking the single-molecule θ temperature to the critical solution temperature is possible for homotypic LLPS when the driving forces for the single-molecule coil-to-globule transitions are similar to those stabilising the phase transition. Therefore, to validate our results, we calculate the θ temperature firstly for WT-FUS and two of our evolved variants by performing all-atom Monte Carlo simulations in implicit solvent using the ABSINTH framework,[36,49] as implemented in the CAMPARI code.[50] The advantage of doing this is that ABSINTH is currently one of the most reliable modelling approaches to produce experimentally consistent conformational ensembles of IDRs[100–103] and to predict IDR critical solution parameters in agreement with experiment.[22,51] The success of ABSINTH is anchored in its extensive experimental validation and refinement, and use of experimentally derived reference solvation free energies. Reassuringly, ABSINTH ranks both our FUS variants and our LAF1 variants from low to high θ temperatures in the same order as the coarse-grained models.

For hnRNPA1, we can not only compute the relative ordering of θ temperatures with ABSINTH, but we can also compute

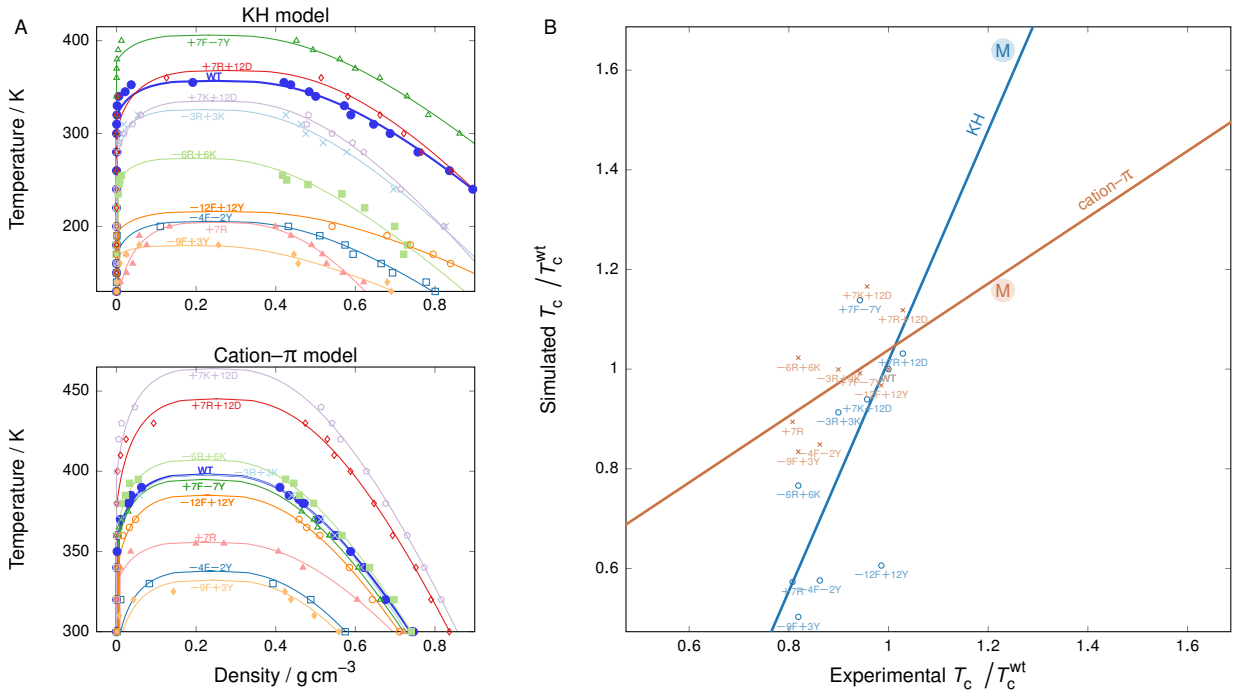

FIG. 7. **Comparison of model predictions with experiment.** In (A), we show phase diagrams obtained with two coarse-grained models for a variety of modifications of the hnRNPA1 sequence. Symbols correspond to simulation results. Solid lines are obtained by the fitting procedure described in SI Sec. S1.2. In (B), we show the correlation plot between simulation and experimental data, alongside a linear fit to the data points. The adjusted squared sum of residuals is $R^2 = 0.96$ for the KH model and $R^2 = 0.99$ for the cation–π model, with p values of 0.0007 and 0.012, respectively. The Pearson correlation coefficients are 0.66 (KH) and 0.64 (cation–π). The point labelled 'M', which was not included when computing the linear fit or the correlation coefficients, corresponds to the sequence with the largest critical point obtained in our genetic-algorithm run (see SI Sec. S9); since there are no experimental data available for this specific sequence, we have estimated the experimental critical temperature using the ABSINTH model.

approximate critical points for the wild-type hnRNPA1 and some of its analogues for which experimental results are available.[23] We have computed these for the wild type and the +7R and +7F−7Y analogue, and these estimates agree very well with the experimental critical points determined in Ref. 23: for WT-hnRNPA1, the experimental critical temperature is ∼348 K and the ABSINTH estimate is ∼345 K; for the +7R analogue, they are ∼280 K and ∼275 K, and for the +7F−7Y analogue, they are ∼328 K and ∼325 K. We have also determined the analogous result for the sequence with the largest critical point obtained in the genetic-algorithm run (see SI Sec. S9), and we show it alongside the experimental results in Fig. 7(B). For both coarse-grained potentials we compare, this point falls very close indeed to the prediction of the linear fit from experimental results, suggesting that the simple coarse-grained potentials are sufficiently powerful to obtain qualitative insight into the phase behaviour of intrinsically disordered proteins.

III. DISCUSSION

In this work, we have proposed an efficient computational method to evolve naturally occurring phase-separating protein sequences. Evolving such sequences can provide insight into which sequence features drive LLPS, both when the proteins are in pure form and when they form part of a multi-component mixture; moreover, our approach could also be extended to design experimentally testable amino-acid sequence mutations which either inhibit or promote the LLPS of protein solutions. Our approach combines state-of-the-art molecular simulations of protein condensates, where each protein is described at the single-amino-acid resolution, with a genetic algorithm grounded in a new fitness function – the difference in composition densities of the protein-poor and protein-rich phases at constant volume – which is both a good proxy for the critical solution temperature and computationally far more tractable to obtain than the critical temperature itself. We have shown that such a simple and computationally inexpensive fitness function is sufficient to evolve the amino-acid sequences of naturally occurring proteins and to shift their phase behaviour in the direction we choose. Moreover, by analysing the effects of small changes to naturally occurring amino-acid sequences, we can draw conclusions about the molecular origins of the local gradient in amino-acid sequence space, adding to conventional analyses of driving forces which usually focus on binding energies. Indeed, an important finding of our work is that we have demonstrated how a genetic-algorithm framework that can alter the LLPS behaviour of proteins also enables us to probe the gradient in amino-acid sequence space directly. This can help us both to extend interpretations of why proteins that drive the formation of condensates might have evolved as they have and to gain greater control over intracellular LLPS.

We have coupled our genetic-algorithm approach to the coarse-grained model of Dignon *et al.*,[35] which is one of the best simple models currently available for probing the phase behaviour of protein solutions. This model has been validated against the single-molecule experimental radii of gyration of a wide range of IDRs,[35] is residue-specific, has been shown to reproduce well the experimental phase behaviour of various proteins under different conditions,[21,39,47,104,105] and is computationally sufficiently inexpensive that it affords the determination of bulk LLPS properties for many sequences. Furthermore, the model accounts for key physicochemical aspects that determine the phase behaviour of proteins, such as the charge, size, relative hydrophobicity and flexibility of amino acids. However, as is the case with any coarse-grained model,

it is still approximate and averages out other effects, in this case especially the specific contribution of π–π interactions, polarisation effects that give rise to cation–π interactions and the explicit role of water and ions in solution. To benchmark the robustness of the protein model, we have therefore repeated our simulations with another coarse-grained potential that can account for cation–π interactions[58] and protein–RNA interactions.[59] The two models predict the same behaviour on evolution, as we discuss in SI Sec. S6, which suggests that the trends we have outlined are not especially sensitive to the choice of model. The evolution of protein amino-acid sequences is a computationally expensive process, and is only made feasible by the choice of a suitably coarse-grained potential. However, it is not immediately clear that such simplified models, which were largely validated against single-molecule experiments, are predictive in the context of LLPS. We have therefore validated the model predictions against experimental results[23] and an all-atom potential;[49,50] we discuss this validation in detail in SI Sec. S10. Despite the fact that the models' predictions are not quantitative, the trends in critical temperature predicted by the models we have used correlate well with experimental results, suggesting that such simple coarse-grained potentials are sufficiently powerful to obtain qualitative insight into the physical driving forces governing the phase behaviour of intrinsically disordered proteins. Moreover, the genetic algorithm we have introduced can of course straightforwardly be used with protein models of higher resolution and accuracy as they are developed, provided that sufficient computational resources are available.

Although the fine details of the phase behaviour we have observed may be model-specific, we have nevertheless shown two distinct driving regimes for enhancing or inhibiting LLPS to exist, namely hydrophobicity – including both the strength and range of relevant interactions – and charge patterning. In sequences such as the PLD of FUS, only one may be in operation, whilst in others, such as hnRNPA1-IDR, they may co-evolve, implying that both driving forces can contribute to LLPS simultaneously. Furthermore, in the hydrophobic driving regime in the case of the PLD of FUS, we have shown that there is a patterning length scale of 2–3 amino acids, which one can interpret in the context of the stickers-and-spacers model of proteins. Intriguingly, we have shown that although LAF1-IDR is charge-rich, charge patterning does not appear to co-evolve with hydrophobicity. In all cases studied, LLPS is facilitated by an increase in the mean size of the amino-acid residues of the proteins, which results in a more structured protein-rich phase, which in turn can favour condensate formation. It would be especially interesting to investigate in future work whether such a driving force for phase separation is more universal than might previously have been thought. Finally, we have demonstrated that the genetic-algorithm approach is successful at evolving sequences in the presence of other species in the medium; here, we have focussed on simple RNA molecules as a proof of concept. Significant changes to the gradient in sequence space are observed when another species is introduced into the system, indicating that our method is also suitable for investigating the co-evolution of proteins and for studying biologically relevant mixtures of different species. Since the effect of sequence modifications on the phase behaviour of many-component mixtures is much less intuitive to predict manually than it is in one-component systems, the ability to guide phase behaviour algorithmically is especially attractive.

In summary, we have presented a powerful framework for systematically modulating the LLPS of proteins by evolving their amino-acid sequences. We have shown that the approach is able to provide direct insight into the nature of LLPS in protein solutions, demonstrating both which fundamental driving forces are in operation as well as providing specific guidance into the kinds of mutation that may help promote or inhibit LLPS in practical applications. Recently, several databases of proteins exhibiting LLPS have been assembled,[106–108] providing an excellent starting point for determining and contrasting the driving forces governing phase separation in very different systems. We have already drawn useful conclusions from the application of our approach to specific cases, contributing a significant piece of the puzzle towards a fuller understanding of the physical driving forces behind LLPS. As ever more accurate force fields of proteins in solution are developed, this approach promises to be particularly fruitful in furthering our understanding of the regulation of LLPS in biology, as well as representing a first step towards future engineering of phase-separating sequences.

IV. METHODS

In the Supplementary Material, we describe the coarse-grained potential, provide further details about the computational methods used, provide further analysis and additional supporting results, and provide the sequences of the proteins studied.

A. Simulation methods.

We performed molecular dynamics simulations of a coarse-grained implicit-solvent model of proteins[35] in which each amino acid is represented as a bead. Neighbouring amino acids in a protein chain are connected by harmonic springs, while other beads interact with one another with a hydrophobicity-scaled Lennard-Jones (LJ) potential and a Debye–Hückel electrostatic potential. The model is discussed in more detail in SI Sec. S1.1, and the simulation methods in SI Sec. S1.2. We have also used a further coarse-grained model for validation[58] and for protein–RNA simulations.[59]

B. Genetic algorithms.

Genetic algorithms optimise properties of a system in ways inspired by biological adaptation of populations.[109–111] All numerical parameters listed below were chosen to balance the need for high evolution speed due to an expensive fitness function and that sufficient diversity in the population be maintained to avoid premature convergence.

1. We define a chromosome of length n, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, where in our case x_{ij} is an amino acid. A set of N chromosomes defines an initial population $U_0 = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$, where U_t denotes the population of a given round t. The starting population in our case corresponds to mutated versions of the WT $\boldsymbol{x}_{\mathrm{WT}}$ with a certain mutation rate, i.e. the frequency at which an amino acid is exchanged for a random one picked from the natural set of 20. We use a rate of 0.01 in this work. A scalar fitness function $f(\boldsymbol{x})$ denotes the property being optimised. We use the width of the phase diagram at a fixed temperature as a proxy for the critical solution temperature, and define the fitness function as

$$f(\boldsymbol{x}) = \frac{\rho_{\mathrm{l}}(\boldsymbol{x}) - \rho_{\mathrm{v}}(\boldsymbol{x})}{\rho_{\mathrm{l}}(\boldsymbol{x}_{\mathrm{WT}}) - \rho_{\mathrm{v}}(\boldsymbol{x}_{\mathrm{WT}})}\,, \tag{2}$$

where $\rho_{\mathrm{l}}(\boldsymbol{x})$ is the average density of the 'liquid-like' protein-rich phase of sequence $\boldsymbol{x}$ and $\rho_{\mathrm{v}}(\boldsymbol{x})$ is the analogue for the 'vapour-like' protein-poor phase. We refer

to individuals with high fitness value as 'strong' and to those with low fitness value as 'weak'. In our case, $N = 20$. For genetic-algorithm runs in which the target is to reduce the critical solution temperature, we use as the fitness function the reciprocal of $f(\boldsymbol{x})$.

2. In each round t, we choose $N_{\text{par}} = 8$ parents P from the population, $P \subset U_t$. To achieve this, we use tournament selection:[111] We first define N_{par} tournaments T_i. Each tournament is a randomly drawn subset of N_{tour} elements from U_t. The fittest sequence from each tournament becomes one of the parents. The tournament size N_{tour} is therefore a direct scaling parameter governing the selection pressure. For our purposes, we have found that $N_{\text{tour}} = 5$ works well.

3. The parents are randomly divided into pairs $(\boldsymbol{a}, \boldsymbol{b})$ (where $\boldsymbol{a} \in P$ and $\boldsymbol{b} \in P$), and crossed over. Here, we swap sequences after a randomly chosen position $k \in [1, n]$ in the sequences, such that

$$(a_i,\, b_i) = \begin{cases} (a_i,\, b_i) & \text{if } i \leq k, \\ (b_i,\, a_i) & \text{otherwise.} \end{cases} \tag{3}$$

4. Another round of random mutations with the same mutation rate is then performed to cover previously unrepresented areas of sequence space.

5. The result from steps 3 and 4 is a set of children C, whose fitness is then evaluated. Children replace some chromosomes in the population. As fitness functions are relatively expensive to compute for our system, we use weak-population replacement,[109] a greedy algorithm that can achieve rapid population evolution. Sequentially, each child $\boldsymbol{c}_i \in C$ is compared to the weakest individual in the population, $\boldsymbol{x}_{\text{weak}} \in U_t$. If $f(\boldsymbol{c}_i) > f(\boldsymbol{x}_{\text{weak}})$, then $\boldsymbol{c}_i$ replaces $\boldsymbol{x}_{\text{weak}}$. The weakest individual may also be a previously inserted child.

Parallelisation can speed up genetic-algorithm progression.[112] We use a simple master–slave approach since asynchronous schemes are not well suited to small populations; because all simulations are run for the same amount of wall-clock time, the overhead of the simple genetic-algorithm parallelisation employed is small compared to the duration of individual simulations.

V. ACKNOWLEDGEMENTS

We thank Jeetain Mittal and Gregory L. Dignon for their invaluable help with implementing their sequence-dependent coarse-grained protein model in LAMMPS. We acknowledge DiRAC funding from the Science and Technology Facilities Council. This project has received funding from the European Research Council (ERC) under the European Union Horizon 2020 research and innovation programme (grant agreement 803326). R. C.-G. is an Advanced Research Fellow of the Winton Programme for the Physics of Sustainability. A. G. is funded by an EPSRC studentship (EP/N509620/1). This work was performed using resources provided by the Cambridge Tier-2 system operated by the University of Cambridge Research Computing Service funded by the EPSRC Tier-2 capital grant EP/P020259/1.

VI. AUTHOR CONTRIBUTIONS

SML, RCG and AR designed the research. SML and AR performed the research. AG assisted with the computational implementation of the coarse-grained protein model. SML, RCG and AR analysed the results and wrote the paper.

SUPPLEMENTARY INFORMATION

S1 COMPUTATIONAL DETAILS

S1.1 Coarse-grained model of proteins

The coarse-grained potential of proteins introduced by Dignon and co-workers[35] is based on an amino-acid level description of protein chains. Each amino acid is represented by a bead. For amino acids that are covalently bonded in the protein of interest, their beads are connected by a harmonic spring,

$$\Phi_{\text{harm}} = \frac{1}{2}k(r - r_0)^2, \tag{S1}$$

where $k = 19.2\,\text{kcal mol}^{-1}\,\text{Å}^{-2}$ is the spring constant and $r_0 = 3.81\,\text{Å}$ is the equilibrium bond length. Beads interact with one another through an Ashbaugh–Hatch modulated[113] hydrophobicity-scaled (12,6)-Lennard-Jones (LJ) potential,

$$\Phi(r)_{ij} = \begin{cases} \Phi_{\text{LJ}}(r) + (1 - \lambda_{ij})\varepsilon_{ij} & \text{if } r < 2^{1/6}\sigma_{ij}, \\ \lambda_{ij}\Phi_{\text{LJ}}(r) & \text{otherwise,} \end{cases} \tag{S2}$$

where r is the interparticle distance, and

$$\Phi_{\text{LJ}}(r) = 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r}\right)^{12} - \left(\frac{\sigma_{ij}}{r}\right)^{6}\right], \tag{S3}$$

where ε_{ij} is the minimum of the LJ potential, σ_{ij} is the LJ diameter and λ_{ij} is a hydrophobicity scaling parameter. In each case, i and j correspond to the amino acid types of the two particles. Residues that carry a charge (see Table I) also interact with a Debye-screened[114] electrostatic potential,

$$\Phi_{\text{coul}}(r) = \frac{q_i q_j}{4\pi D r}\mathrm{e}^{-r/\kappa}, \tag{S4}$$

where D is the permittivity $D = 80\,\varepsilon_0$ (where ε_0 is the electric constant), κ is the screening length ($\kappa = 1\,\text{nm}$, corresponding to an ionic strength of 100 mM) and q_i and q_j are the charges of the amino acids.

In the work of Dignon and co-workers,[35] two possibilities for assigning λ_{ij}, ε_{ij} and σ_{ij} are presented. Here, we use the variant based on the Kim–Hummer (KH) model,[115] which has been brought into the form of equation (S2) with parameters

$$\varepsilon_{ij} = |\alpha(\varepsilon_{\text{MJ}} - \varepsilon_0)| \quad \text{and} \quad \lambda_{ij} = \begin{cases} 1 & \text{if } \varepsilon_{\text{MJ}} \leq \varepsilon_0, \\ -1 & \text{otherwise,} \end{cases} \tag{S5}$$

where ε_{MJ} is the Miyazawa–Jerningan empirical contact potential,[116] and $\alpha = 0.228$ and $\varepsilon_0 = 1\,\text{kcal mol}^{-1}$ are benchmarked on experimental radii of gyration. The σ_{ij} parameters are arithmetic means of effective Van der Waals radii of the amino acids, $\sigma_{ij} = (\sigma_i + \sigma_j)/2$, which are given in Table I.

A version of this model with a new parameter set[38] was recently introduced, featuring a temperature dependence of

TABLE I. **Table of amino acids.** The 20 naturally occurring amino acids with their one- and three-letter codes, alongside their charges. Amino acids marked with a '⋆' are aromatic. The last columns give the λ- and σ-parameters which define the hydrophobicity scale.

Full name	Code		Charge	λ	σ/Å
Arginine	Arg	R	+	0.0	6.56
Aspartate	Asp	D	−	0.378	5.58
Asparagine	Asn	N	0	0.432	5.68
Glutamate	Glu	E	−	0.459	5.92
Lysine	Lys	K	+	0.514	6.36
Histidine	His	H	+	0.514	6.08
Glutamine	Gln	Q	0	0.514	6.02
Cysteine	Cys	C	0	0.595	5.48
Serine	Ser	S	0	0.595	5.18
Glycine	Gly	G	0	0.649	4.50
Threonine	Thr	T	0	0.676	5.62
Alanine	Ala	A	0	0.730	5.04
Methionine	Met	M	0	0.838	6.18
⋆ Tyrosine	Tyr	Y	0	0.865	6.46
Valine	Val	V	0	0.891	5.86
⋆ Tryptophan	Trp	W	0	0.946	6.78
Leucine	Leu	L	0	0.973	6.18
Isoleucine	Ile	I	0	0.973	6.18
Proline	Pro	P	0	1.0	5.56
⋆ Phenylalanine	Phe	F	0	1.0	6.36

the λ_{ij} parameters to match experimental data more closely. We have not employed this potential here, as we are primarily interested in relative shifts of T_c and not its absolute value.

As an alternative to the Kim–Hummer model, varying $\lambda_{ij} = (\lambda_i + \lambda_j)/2$ parameters in an amino-acid specific way is also possible, while taking ε_{ij} to be constant. This has been shown to yield comparable results.[35] Whilst we do not employ this model here, we use these λ_i values (Table I) of the amino acids to quantify their hydrophobicity.

S1.2 Determining phase coexistence: simulation details

We studied LLPS with direct-coexistence molecular dynamics simulations[117] in which a high- and a low-density phase coexist. Molecules can be exchanged between the two phases, allowing the densities to equilibrate across the interface [Fig. S1(A))]. In the limit of sufficiently large systems, where the interface becomes negligible compared to the bulk of each phase, this approach allows us to determine the densities of the two compositional phases. Direct-coexistence simulations provide an especially simple method of determining phase equilibria, particularly with liquid-like phases considered here. Since phase separation above the spinodal is a nucleation-initiated process, hysteresis may be a problem, and direct-coexistence simulations may require a careful calibration of the interface at the start of a simulation.[75] However, with the kinds of coarse-grained potentials we are using, phase transitions are facile and an interface forms readily. Direct-coexistence simulations have therefore routinely been employed in computational studies of LLPS with such models.[118]

In order to construct a phase diagram, we perform direct-coexistence simulations at a number of temperatures [Fig. S1(A)]. We determine the densities of the coexisting phases by binning particles along the z axis and finding a least-squares best fit to two constant densities – with an interface of a system-specific width – to identify the low- and the high-density phases as well as the interfacial region [Fig. S1(B))]. Finally,

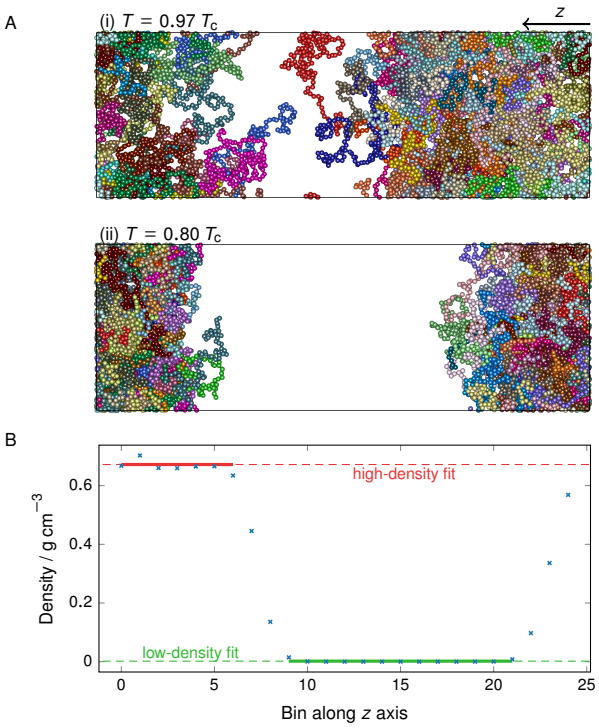

FIG. S1. **Direct-coexistence simulations.** (A) Orthographic projections of simulation boxes of FUS at two temperatures below T_c, as indicated, where different colours represent different protein chains. The box is periodic in all directions. (B) Example of a fitted density profile for FUS at $T = 0.8\,T_c$. The simulation box along the z direction is split into 25 bins and the average density is computed within each bin, indicated by blue crosses. Points from the regions where solid lines are shown were used to compute a fit to a constant for each of the high- and the low-density phases, taking into account an interface of finite thickness.

the densities we extract are plotted on a temperature–density phase diagram, as for example in Fig. S2. To interpolate the data points close to the critical temperature, we use an empirical fit[31] which, although it does not capture the behaviour of the system very well at low temperatures, is sufficient to find the approximate critical temperature and density. In particular, we fit simultaneously to

$$(\rho_{\text{high}}(T) - \rho_{\text{low}}(T))^{3.06} = d\,(1 - T/T_c) \qquad \text{and} \qquad \text{(S6)}$$

$$\rho_{\text{high}}(T) + \rho_{\text{low}}(T) = 2\rho_c + 2s_2(T_c - T), \qquad \text{(S7)}$$

where $\rho_{\text{high}}(T)$, $\rho_{\text{low}}(T)$ and ρ_c are the densities of the high-density and low-density phases and the critical density, respectively; T_c is the critical temperature and d and s_2 are fitting parameters. This works by numerically finding a best fit for the four constants – ρ_c, T_c, d and s_2 – from equations constrained by all pairs of $\rho_{\text{high}}(T)$ and $\rho_{\text{low}}(T)$ deemed by inspection of the curvature of the observed data series to lie below T_c.

We note that in direct-coexistence simulations, above the critical temperature, the system forms a single supercritical fluid; the density fitting outlined above, which assumes two distinct phases have formed, will thus produce non-physical results reflecting the natural density fluctuations of the system; any densities so determined do not correspond to coexisting phases. Nevertheless, such spurious points help us to ascertain that we have already crossed the critical point, forming a characteristic 'protrusion' of the coexistence region towards higher temperatures. The presence of such features at the appropriate point can thus provide an additional check that we have determined the critical temperature correctly, and so we

have included them in phase diagrams as greyed-out points for reference.

We performed molecular dynamics simulations with the LAMMPS simulation package,[119] using a velocity Verlet integrator with a time step of $\delta t = 10\,\mathrm{fs}$. Direct-coexistence simulations were run in the canonical ensemble with a Langevin thermostat[120] with a damping time of $10^4\,\delta t$. We used a tetragonal simulation box with periodic boundary conditions, with typical dimensions 134.4 Å × 134.4 Å × 403.2 Å with 64 chains for FUS and hnRNPA1, and 267.2 Å × 267.2 Å × 1336.2 Å with 512 chains for LAF1. We have verified that for typical systems, these numbers of copies of the polymer chain were sufficient to ensure that finite-size effects did not dominate the system's bulk behaviour; to this end, we simulated systems ∼30 % smaller and verified that the mean densities computed for these smaller systems were sufficiently similar that the estimated critical temperature fell within 5 K of the original one, which is within typical error bars of temperature measurements with Langevin thermostats.

S2 HYDROPHOBIC SEQUENCE OPTIMUM

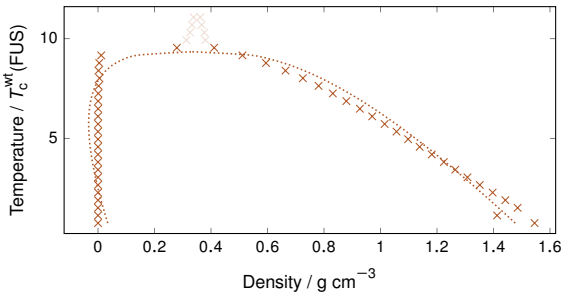

FIG. S2. **Hydrophobic sequence optimum.** Phase diagram of the $(\mathrm{Phe})_{163}$ sequence, the same length as FUS-WT. The critical temperature is approximately nine times that of FUS-WT. The dotted line is a fit, and greyed-out points lie above the critical point, as detailed in Section S1.2.

Figure S2 shows the phase diagram of a possible optimum of hydrophobicity-driven LLPS within the CG model. While assessing an optimum of the charge-driven case is difficult, it possible to make a guess about the scaled LJ potential by inspection of the model parameters. The highest ε_{ij} value is for a Phe–Phe interaction, which is also longer-ranged than for most other amino acids. The phase diagram shown illustrates how large the changes in critical temperature can be, but it also demonstrates that trivially optimising the critical temperature results in a biologically uninteresting sequence.

S3 DUMMY GENETIC ALGORITHM

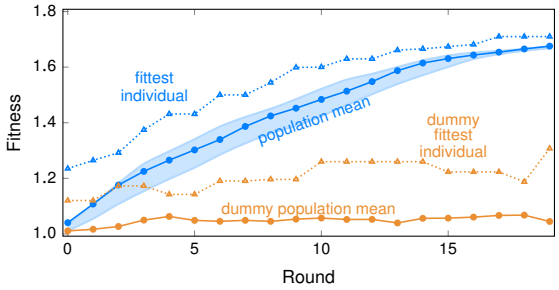

FIG. S3. **Dummy genetic algorithm.** A dummy genetic-algorithm progression (orange), contrasted with the driven evolution (blue) of FUS phase separation. We remove any driving force by randomising the selection of parents and replacement in the population, but keeping all mutation steps.

Figure S3 shows the genetic algorithm progression for FUS with enhanced phase separation, contrasted with a dummy genetic algorithm without any selection pressure. By setting the tournament size $N_{\mathrm{tour}} = 1$, we remove any favouring of fit parents for the generation of children, and instead of performing weak-population replacement, we replace a random individual within the population with an acceptance probability of 0.5. However, we leave all mutagenesis steps intact, thus setting up a control for how the random mutagenesis itself affects FUS phase separation. In Fig. S3, we show that such a dummy genetic algortihm exhibits very little evolutionary driving force compared to our original implementation. The observed increases in phase-separating ability are therefore not due to mere random alterations of the sequences, but are directly driven by our genetic algorithm.

S4 REDUCING THE CRITICAL TEMPERATURE

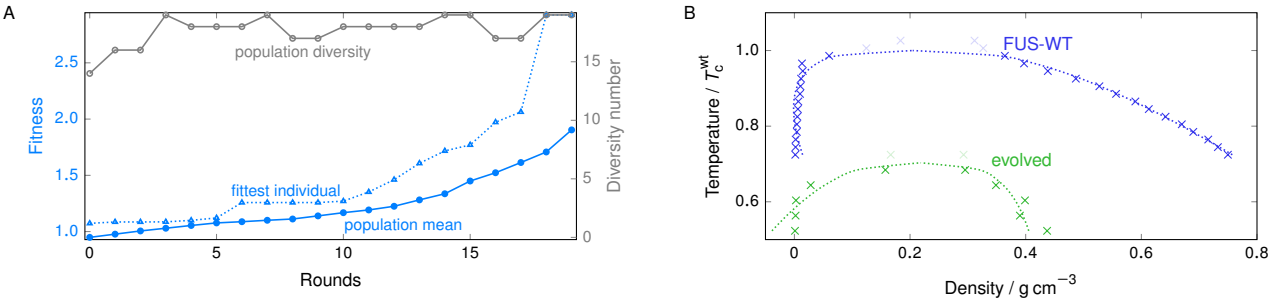

FIG. S4. **Reducing the critical temperature.** (A) Typical GA progression for FUS where the fitness function *reduces* the upper critical solution temperature. The fitness function [defined as the reciprocal of Eq. (2) of the main manuscript] increases by ∼90 % over 20 rounds. The fittest individual can be considerably fitter than the mean. The population diversity, i.e. the number of distinct sequences present in the overall population of 20, is generally very high. (B) Comparison of representative phase diagrams before and after genetic-algorithm runs, confirming that the fitness function choice was suitable. Dotted lines are fits, and greyed-out points lie above the critical point, as detailed in Section S1.2.

S5 CHUNK-SHUFFLING EXAMPLE RUNS

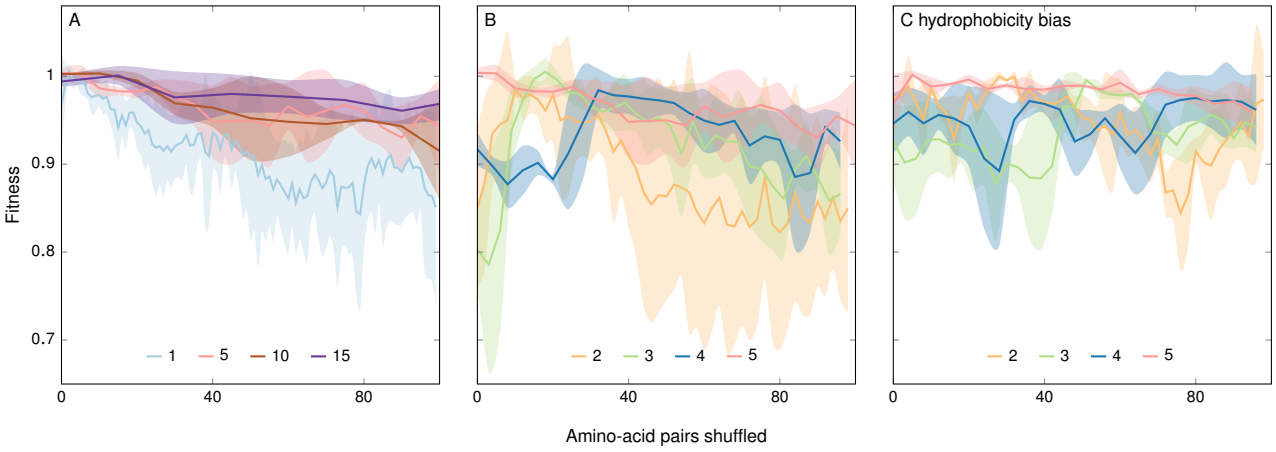

FIG. S5. **Chunk-shuffling example runs.** (A) Chunk shuffling at chunk lengths 1, 5, 10 and 15. Chunk length 1 lies significantly below the other curves. Shaded areas are standard deviations from 3 (6 for chunk length 1 and 5) shuffling runs. (B) Chunk shuffling with focus on chunk lengths around a length scale of 2–3 amino acids. Shaded areas are standard deviations from 3 (6 for chunk lengths 2 and 5) shuffling runs. (C) Hydrophobicity-biassed chunk shuffling, only allowing exchanges between the top 30 % chunks by hydrophobicity. Shaded areas are standard deviations from 3 shuffling runs.

S6 EVOLUTION WITH THE CATION–π MODEL

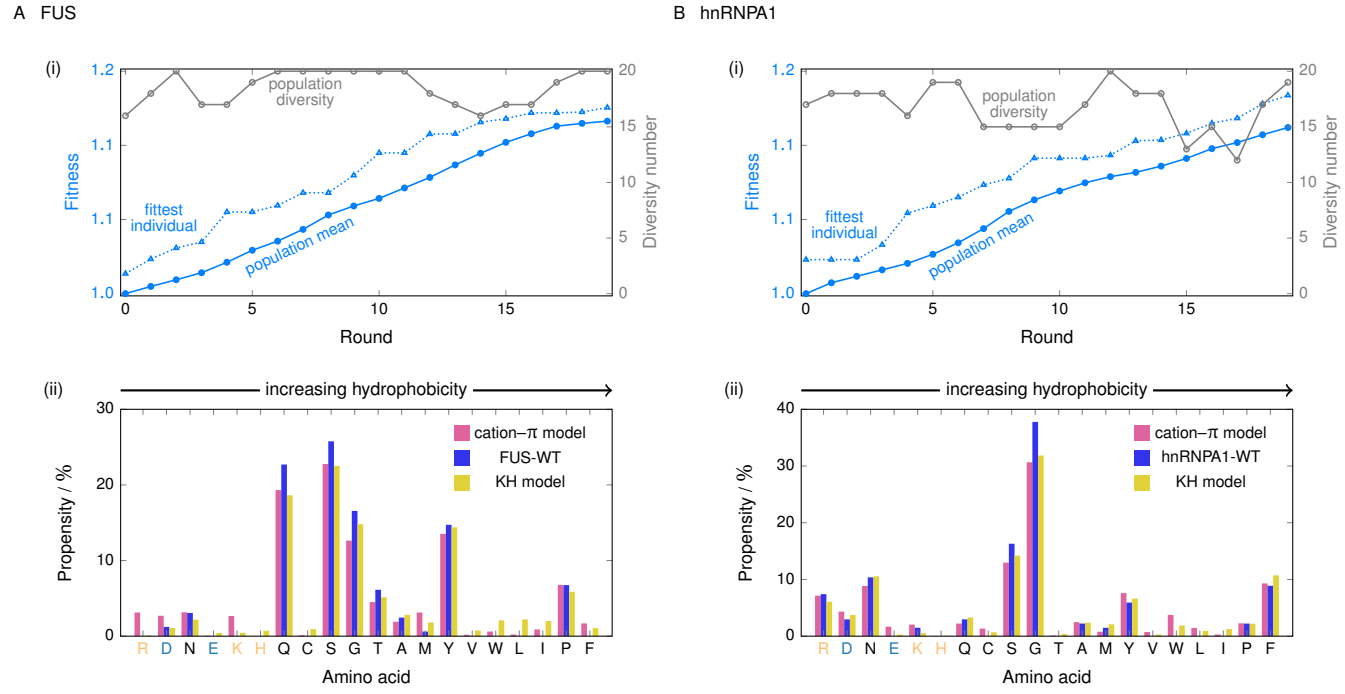

FIG. S6. **Evolution with the cation–π model.** Results from the application of the genetic algorithm to (A) FUS and (B) hnRNPA1 using the cation–π model. In each case, panel (i) shows the change in fitness function and the population diversity as a function of the genetic-algorithm round, while panel (ii) shows a comparison of the evolved amino-acid propensities for the two sequences when evolved with the model used elsewhere ('KH') and the cation–π reparameterisation, with the wild type shown for reference. Amino acids are plotted in order of increasing hydrophobicity [see Table I]. Positively charged amino acids are indicated in light orange and negatively ones in light blue.

In order to determine how sensitive the results of the predictions of the genetic-algorithm runs are to the choice of model, we have re-run evolution simulations of FUS and hnRNPA1 with a reparameterised coarse-grained potential of Das *et al.*,[58] which we refer to as the cation–π model. The approach is similar to the KH model we introduced in Section S1.1; each amino acid in the sequence is represented by a bead and the primary interaction is still of a Lennard-Jones type coupled with Debye–Hückel ionic terms, but the cation–π model uses the hydrophobicity scale directly, i.e. by changing λ rather than ε values, and additionally accounts for cation–π interactions for arginine and leucine with phenylalanine, tryptophan and

tyrosine residues with an additional Lennard-Jones interaction. We use 'scheme (i)' from the paper of Das *et al.*[58] in our implementation, and identical parameters for the harmonic spring bond length and spring constant.

We show the results of the evolution of FUS and hnRNPA1 with our genetic algorithm in Fig. S6. These results are largely comparable with the results shown in Fig. 1(B) of the main text for FUS and Fig. 4(A) of the main text for hnRNPA1. Although the change in the fitness function as a function of the genetic-algorithm run is slower than that for the KH-model analogue, this is not unexpected, since the width of the phase diagram, which is how the fitness function is determined, varies less with the critical temperature in the cation–π model [as we show in Fig. 7 of the main text]. While some model-specific differences are expected, we can see from Fig. S6(B) that these amount mainly to a small increase in favourability of cations, which is expected from the goal of the reparameterisation of the cation–π potential, and a slight narrowing of the broadness of the hydrophobic distribution of residues generated. Although of course these models are ultimately based on the same coarse-graining philosophy, and so the good agreement between their predictions is perhaps not entirely unexpected, these results suggest that the broad trends that we have discussed in the main text are not overly sensitive with respect to the choice of model.

S7 LIQUID CHARACTER OF PHASES

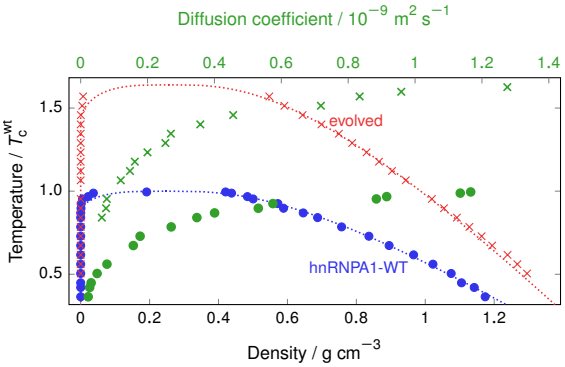

FIG. S7. **Liquid character of phases.** Phase diagram of wild-type hnRNPA1 (in blue) and of the hnRNPA1 analogue obtained at the end of the genetic-algorithm run (in red), alongside diffusion coefficients for the liquid-like phase (green circles for the wild type, green crosses for the evolved sequence, measured along the green abscissa). The densities of the coexisting phases were first determined in direct-coexistence simulations at a range of temperatures. The diffusion coefficient was then computed in a canonical-ensemble simulation at the density corresponding to the liquid-like (high-density) phase. A non-linear relationship exists between the diffusion coefficient and temperature, but there is no indication of a discontinuous change that might indicate a glass transition.

When studying LLPS, it is important to control the liquid character of the observed phases, as proteins can also undergo gelation[55] and glass transitions.[121] While the study of general polymer dynamics in the CG model is not the objective of this work, and dynamical properties of coarse-grained models do not usually faithfully reproduce real dynamics, we nevertheless computed the per-bead diffusion coefficient for our systems to ascertain that a glass transition has not occurred. Ballistic and sub-diffusive regimes may precede the diffusive regime, and in polymer systems, several sub-diffusive regimes can often be observed.[122,123] The diffusion coefficient D can be obtained by the Einstein relation[124] for the mean squared displacement, $\langle \Delta r^2 \rangle = 6Dt$, which holds in the diffusive regime of *long* times t. We have computed diffusion coefficients in this way over a range of temperatures, and show these alongside the phase diagram of hnRNPA1-WT and one of the evolved sequences in Fig. S7. In all cases, the diffusive regime can easily be reached in a brute-force simulation at readily accessible time scales, suggesting that the systems are not dynamically arrested under the conditions of interest. The variation in the diffusion coefficients is almost entirely due to the density; the temperature dependence simply corresponds to the fact that the liquid phase is less dense at a given temperature for the wild-type than it is for the evolved sequence. The diffusion coefficients shown are only qualitative, in the sense that in the potential we use, many degrees of freedom have been coarse-grained away, and the unit of time is not directly comparable to experiment. However, ratios of diffusion coefficients are nevertheless meaningful. The temperature dependence of the diffusion coefficients is non-linear; however, there is no obvious discontinuous change in the diffusion coefficient as a function of temperature in these data even at low temperatures in the liquid phase, which we take as justification that our simulations describe LLPS rather than glass formation, which could complicate our interpretation of the results.

S8 AMINO-ACID SEQUENCES OF PROTEINS STUDIED

We give below the amino-acid sequences of the prion-like IDR of FUS, hnRNPA1-IDR and LAF1-IDR[30] studied in this work, using one-letter codes [Table I] for the amino acids.

[residues 1–163 of UniProt sequence P35637-1]

FUS
MASND YTQQA TQSYG AYPTQ PGQGY SQQSS QPYGQ
QSYSG YSQST DTSGY GQSSY SSYGQ SQNTG YGTQS
TPQGY GSTGG YGSSQ SSQSS YGQQS SYPGY GQQPA
PSSTS GSYGS SSQSS SYGQP QSGSY SQQPS YGGQQ
QSYGQ QQSYN PPQGY GQQNQ YNS

[residues 186–320 of UniProt sequence P09651-2]

hnRNPA1
MASAS SSQRG RSGSG NFGGG RGGGF GGNDN FGRGG
NFSGR GGFGG SRGGG GYGGS GDGYN GFGND GSNFG
GGGSY NDFGN YNNQS SNFGP MKGGN FGGRS SGPYG
GGGQY FAKPR NQGGY GGSSS SSSYG SGRRF

[residues 2–168 of UniProt sequence D0PV95-1]

LAF1
ESNQS NNGGS GNAAL NRGGR YVPPH LRGGD GGAAA
AASAG GDDRR GGAGG GGYRR GGGNS GGGGG GGYDR
GYNDN RDDRD NRGGS GGYGR DRNYE DRGYN GGGGG
GGNRG YNNNR GGGGG GYNRQ DRGDG GSSNF SRGGY
NNRDE GSDNR GSGRS YNNDR RDNGG DG

S9 EVOLVED SEQUENCES

Below, we provide example output sequences, taken from the final populations of the appropriate GA runs as described in this work, using one-letter codes [Table I]. The residues that have changed compared to the initial sequence are highlighted in red. Sequences of full populations as a function of the progression of all our GA runs can be found in the supporting data. In the list below, for the case of FUS, 'max' and 'min' refer to the sequences with the highest and lowest critical temperatures at the end of the genetic-algorithm run when the fitness function was designed to increase and to decrease the width of the phase diagram, respectively.

FUS max
MASFD YLMYA QQSYG AYGTQ PYQIY SQQSP QPYHM
QPYSG YSQST YTSGY GMSSM YPYGQ SQNTG YGTQS
WPLGY GSTGG CGSSQ SSQSS IGQQG SYWGY GQQPA
PSSTS YFYGS SSQSS SWGQK QSGSY SQLPS YGGQQ
YSYGQ QQSYN PHQGY WQQWQ YHS

FUS min
MASNM YPQQA TQSYG AYRTQ PGTGY SKQSS QPYGQ
QSYKG YCGVT GTSGE GQSSY KSYGQ SQNTG SGTQS
KPQGY GSTGG YGSSQ GSKSK PGQQS SYNGI GQQPA
RSSTS GSYGG KSQSS SYGQP QSGSP SQQPS DGGQQ
QSGGQ QQSYN PPQGY GQQQQ YND

hnRNPA1
FASAS SSQRG NSGSG NFPGC TIDGF GKNDN FGNGG
NFSGR GWFGG CRGGP WYGFS GDGYN GFGND GSNFG
YFVSY NDFGN YNEQS SNFDP MRNGN FIGYS SGPYG
GGGQF FARFR IQGGY GGSSS SSSYM SGRRF

LAF1
ESNQS NNGGH GYAAL NRGGR YVPPH LRGGD GGAAA
AASAG GDDRR GGAGG GFYRR GGGNS GNGGG GDYDR
GYNDN RDDRD NRGGS GGYGW PRNYE DRGYN GGGGA
GGNRS YNNNR GGGEV GYNRQ DRGDG GSSNF SRGDY
NNRDE GSDNR GSGRS YNNDR RDNGG DG

S10 SEQUENCES USED IN EXPERIMENTAL VALIDATION

For benchmarking the predictions of the models used, we have considered the following sequences of hnRNPA1 IDP variants, using the nomenclature of Bremer and co-workers.[23] The names of the variants correspond to the one-letter amino-acid codes of residues that replace those in the wild-type; they are highlighted in red below.

WT
MASAS SSQRG RSGSG NFGGG RGGGF GGNDN FGRGG
NFSGR GGFGG SRGGG GYGGS GDGYN GFGND GSNFG
GGGSY NDFGN YNNQS SNFGP MKGGN FGGRS SGPYG
GGGQY FAKPR NQGGY GGSSS SSSYG SGRRF

−3R+3K
MASAS SSQRG KSGSG NFGGG RGGGF GGNDN FGRGG
NFSGR GGFGG SKGGG GYGGS GDGYN GFGND GSNFG
GGGSY NDFGN YNNQS SNFGP MKGGN FGGRS SGGSG
GGGQY FAKPR NQGGY GGSSS SSSYG SGRKF

−4F−2Y
MASAS SSQRG RSGSG NSGGG RGGGF GGNDN FGRGG
NSSGR GGFGG SRGGG GYGGS GDGYN GFGND GSNSG
GGGSS NDFGN YNNQS SNFGP MKGGN FGGRS SGGSG
GGGQY SAKPR NQGGY GGSSS SSSSG SGRRF

−6R+6K
MASAS SSQKG KSGSG NFGGG RGGGF GGNDN FGKGG
NFSGR GGFGG SKGGG GYGGS GDGYN GFGND GSNFG
GGGSY NDFGN YNNQS SNFGP MKGGN FGGKS SGGSG
GGGQY FAKPR NQGGY GGSSS SSSYG SGRKF

+7F−7Y
MASAS SSQRG RSGSG NFGGG RGGGF GGNDN FGRGG
NFSGR GGFGG SRGGG GFGGS GDGFN GFGND GSNFG
GGGSF NDFGN FNNQS SNFGP MKGGN FGGRS SGGSG
GGGQF FAKPR NQGGF GGSSS SSSFG SGRRF

+7K+12D
MASAD SSQRD RDDKG NFGDG RGGGF GGNDN FGRGG
NFSDR GGFGG SRGDG KYGGD GDKYN GFGND GKNFG
GGGSY NDFGN YNNQS SNFDP MKGGN FKDRS SGPYD
KGGQY FAKPR NQGGY GGSSS SKSYG SDRRF

+7R
MASAS SSQRG RSGRG NFGGG RGGGF GGNDN FGRGG
NFSGR GGFGG SRGGG RYGGS GDRYN GFGND GRNFG
GGGSY NDFGN YNNQS SNFGP MKGGN FRGRS SGPYG
RGGQY FAKPR NQGGY GGSSS SRSYG SGRRF

+7R+12D
MASAD SSQRD RDDRG NFGDG RGGGF GGNDN FGRGG
NFSDR GGFGG SRGDG RYGGD GDRYN GFGND GRNFG
GGGSY NDFGN YNNQS SNFDP MKGGN FRDRS SGPYD
RGGQY FAKPR NQGGY GGSSS SRSYG SDRRF

−9F+3Y
MASAS SSQRG RSGSG NFGGG RGGGY GGNDN GGRGG
NYSGR GGFGG SRGGG GYGGS GDGYN GGGND GSNYG
GGGSY NDSGN GNNQS SNFGP MKGGN YGGRS SGGSG
GGGQY GAKPR NQGGY GGSSS SSSYG SGRRS

−12F+12Y
MASAS SSQRG RSGSG NYGGG RGGGY GGNDN YGRGG
NYSGR GGYGG SRGGG GYGGS GDGYN GYGND GSNYG
GGGSY NDYGN YNNQS SNYGP MKGGN YGGRS SGGSG
GGGQY YAKPR NQGGY GGSSS SSSYG SGRRY

S11 REFERENCES

[1] Y. Shin and C. P. Brangwynne, "Liquid phase condensation in cell physiology and disease," Science **357**, eaaf4382 (2017).
[2] S. Boeynaems, S. Alberti, N. L. Fawzi, T. Mittag, M. Polymenidou, F. Rousseau, J. Schymkowitz, J. Shorter, B. Wolozin, L. Van Den Bosch, P. Tompa, and M. Fuxreiter, "Protein phase separation: A new phase in cell biology," Trends Cell Biol. **28**, 420 (2018).
[3] A. Molliex, J. Temirov, J. Lee, M. Coughlin, A. P. Kanagaraj, H. J. Kim, T. Mittag, and J. P. Taylor, "Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization," Cell **163**, 123 (2015).
[4] C. P. Brangwynne, C. R. Eckmann, D. S. Courson, A. Rybarska, C. Hoege, J. Gharakhani, F. Jülicher, and A. A. Hyman, "Germline P granules are liquid droplets that localize by controlled dissolution/condensation," Science **324**, 1729 (2009).
[5] A. Putnam, M. Cassani, J. Smith, and G. Seydoux, "A gel phase promotes condensation of liquid P granules in *Caenorhabditis elegans* embryos," Nat. Struct. Mol. Biol. **26**, 220 (2019).
[6] P. Li, S. Banjade, H.-C. Cheng, S. Kim, B. Chen, L. Guo, M. Llaguno, J. V. Hollingsworth, D. S. King, S. F. Banani, P. S. Russo, Q.-X. Jiang, B. T. Nixon, and M. K. Rosen, "Phase transitions in the assembly of multivalent signalling proteins," Nature **483**, 336 (2012).
[7] C. P. Brangwynne, T. J. Mitchison, and A. A. Hyman, "Active liquid-like behavior of nucleoli determines their size and shape in *Xenopus laevis* oocytes," Proc. Natl. Acad. Sci. U. S. A. **108**, 4334 (2011).
[8] A. G. Larson, D. Elnatan, M. M. Keenen, M. J. Trnka, J. B. Johnston, A. L. Burlingame, D. A. Agard, S. Redding, and G. J. Narlikar, "Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin," Nature **547**, 236 (2017).
[9] A. R. Strom, A. V. Emelyanov, M. Mir, D. V. Fyodorov, X. Darzacq, and G. H. Karpen, "Phase separation drives heterochromatin domain formation," Nature **547**, 241 (2017).
[10] S. Sanulli, M. J. Trnka, V. Dharmarajan, R. W. Tibble, B. D. Pascal, A. L. Burlingame, P. R. Griffin, J. D. Gross, and G. J. Narlikar, "HP1 reshapes nucleosome core to promote phase separation of heterochromatin," Nature **575**, 390 (2019).
[11] D. Hnisz, K. Shrinivas, R. A. Young, A. K. Chakraborty, and P. A. Sharp, "A phase separation model for transcriptional control," Cell **169**, 13 (2017).
[12] A. Klosin, F. Oltsch, T. Harmon, A. Honigmann, F. Jülicher, A. A. Hyman, and C. Zechner, "Phase separation provides a mechanism to reduce noise in cells," Science **367**, 464 (2020).
[13] O. A. Saleh, B.-j. Jeon, and T. Liedl, "Enzymatic degradation of liquid droplets of DNA is modulated near the phase boundary," Proc. Natl. Acad. Sci. U. S. A. **117**, 16160 (2020).
[14] F. G. Quiroz, V. F. Fiore, J. Levorse, L. Polak, E. Wong, H. A. Pasolli, and E. Fuchs, "Liquid-liquid phase separation drives skin barrier formation," Science **367**, eaax9554 (2020).
[15] S. Alberti and D. Dormann, "Liquid–liquid phase separation in disease," Annu. Rev. Genet. **53**, 171 (2019).
[16] I. A. Klein, A. Boija, L. K. Afeyan, S. W. Hawken, M. Fan, A. Dall'Agnese, O. Oksuz, J. E. Henninger, K. Shrinivas, B. R. Sabari, I. Sagi, V. E. Clark, J. M. Platt, M. Kar, P. M. McCall, A. V. Zamudio, J. C. Manteiga, E. L. Coffey, C. H. Li, N. M. Hannett, Y. E. Guo, T.-M. Decker, T. I. Lee, T. Zhang, J.-K. Weng, D. J. Taatjes, A. Chakraborty, P. A. Sharp, Y. T. Chang, A. A. Hyman, N. S. Gray, and R. A. Young, "Partitioning of cancer therapeutics in nuclear condensates," Science **368**, 1386 (2020).

[17] P. J. Flory, *Principles of Polymer Chemistry* (Cornell University Press, 1953).

[18] Y.-H. Lin, J. D. Forman-Kay, and H. S. Chan, "Theories for sequence-dependent phase behaviors of biomolecular condensates," Biochemistry **57**, 2499 (2018).

[19] A. A. Hyman, C. A. Weber, and F. Jülicher, "Liquid-liquid phase separation in biology," Annu. Rev. Cell Develop. Biol. **30**, 39 (2014).

[20] C. P. Brangwynne, P. Tompa, and R. V. Pappu, "Polymer physics of intracellular phase transitions," Nat. Phys. **11**, 899 (2015).

[21] G. Krainer, T. J. Welsh, J. A. Joseph, J. R. Espinosa, E. d. Csilléry, A. Sridhar, Z. Toprakcioglu, G. Gudiškytė, M. A. Czekalska, W. E. Arter, P. S. George-Hyslop, R. Collepardo-Guevara, S. Alberti, and T. P. Knowles, "Reentrant liquid condensate phase of proteins is stabilized by hydrophobic and non-ionic interactions," Nat. Commun. **12**, 1085 (2021).

[22] E. W. Martin, A. S. Holehouse, I. Peran, M. Farag, J. J. Incicco, A. Bremer, C. R. Grace, A. Soranno, R. V. Pappu, and T. Mittag, "Valence and patterning of aromatic residues determine the phase behavior of prion-like domains," Science **367**, 694 (2020).

[23] A. Bremer, M. Farag, W. M. Borcherds, I. Peran, E. W. Martin, R. V. Pappu, and T. Mittag, "Deciphering how naturally occurring sequence features impact the phase behaviors of disordered prion-like domains," bioRxiv (2021), 10.1101/2021.01.01.425046.

[24] S. F. Banani, A. M. Rice, W. B. Peeples, Y. Lin, S. Jain, R. Parker, and M. K. Rosen, "Compositional control of phase-separated cellular bodies," Cell **166**, 651 (2016).

[25] J. A. Ditlev, L. B. Case, and M. K. Rosen, "Who's in and who's out—Compositional control of biomolecular condensates," J. Mol. Biol. **430**, 4666 (2018).

[26] J. R. Espinosa, J. A. Joseph, I. Sanchez-Burgos, A. Garaizar, D. Frenkel, and R. Collepardo-Guevara, "Liquid network connectivity regulates the stability and composition of biomolecular condensates with many components," Proc. Natl. Acad. Sci. U. S. A. **117**, 13238 (2020).

[27] J. S. Andersen, Y. W. Lam, A. K. L. Leung, S.-E. Ong, C. E. Lyon, A. I. Lamond, and M. Mann, "Nucleolar proteome dynamics," Nature **433**, 77 (2005).

[28] S. Jain, J. R. Wheeler, R. W. Walters, A. Agrawal, A. Barsic, and R. Parker, "ATPase-modulated stress granules contain a diverse proteome and substructure," Cell **164**, 487 (2016).

[29] A. L. Darling, Y. Liu, C. J. Oldfield, and V. N. Uversky, "Intrinsically disordered proteome of human membrane-less organelles," Proteomics **18**, 1700193 (2018).

[30] Q. Li, X. Peng, Y. Li, W. Tang, J. Zhu, J. Huang, Y. Qi, and Z. Zhang, "LLPSDB: a database of proteins undergoing liquid–liquid phase separation in vitro," Nucleic Acids Res. **48**, D320 (2019).

[31] J. R. Espinosa, A. Garaizar, C. Vega, D. Frenkel, and R. Collepardo-Guevara, "Breakdown of the law of rectilinear diameter and related surprises in the liquid-vapor coexistence in systems of patchy particles," J. Chem. Phys. **150**, 224510 (2019).

[32] F. G. Quiroz and A. Chilkoti, "Sequence heuristics to encode phase behaviour in intrinsically disordered protein polymers," Nat. Mater. **14**, 1164 (2015).

[33] J. Wang, J.-M. Choi, A. S. Holehouse, H. O. Lee, X. Zhang, M. Jahnel, S. Maharana, R. Lemaitre, A. Pozniakovsky, D. Drechsel, I. Poser, R. V. Pappu, S. Alberti, and A. A. Hyman, "A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins," Cell **174**, 688 (2018).

[34] E. W. Martin and T. Mittag, "Relationship of sequence and phase separation in protein low-complexity regions," Biochemistry **57**, 2478—2487 (2018).

[35] G. L. Dignon, W. Zheng, Y. C. Kim, R. B. Best, and J. Mittal, "Sequence determinants of protein phase behavior from a coarse-grained model," PLoS Comput. Biol. **14**, 1 (2018).

[36] J.-M. Choi, F. Dar, and R. V. Pappu, "LASSI: A lattice model for simulating phase transitions of multivalent proteins," PLoS Comput. Biol. **15**, e1007028 (2019).

[37] K. M. Ruff, R. V. Pappu, and A. S. Holehouse, "Conformational preferences and phase behavior of intrinsically disordered low complexity sequences: insights from multiscale simulations," Curr. Opin. Struct. Biol. **56**, 1 (2019).

[38] G. L. Dignon, W. Zheng, Y. C. Kim, and J. Mittal, "Temperature-controlled liquid–liquid phase separation of disordered proteins," ACS Cent. Sci. **5**, 821 (2019).

[39] B. S. Schuster, G. L. Dignon, W. S. Tang, F. M. Kelley, A. K. Ranganath, C. N. Jahnke, A. G. Simpkins, R. M. Regy, D. A. Hammer, M. C. Good, and J. Mittal, "Identifying sequence perturbations to an intrinsically disordered protein that determine its phase-separation behavior," Proc. Natl. Acad. Sci. U. S. A. (2020), 10.1073/pnas.2000223117.

[40] M. Heidenreich, J. M. Georgeson, E. Locatelli, L. Rovigatti, S. K. Nandi, A. Steinberg, Y. Nadav, E. Shimoni, S. A. Safran, J. P. K. Doye, and E. D. Levy, "Designer protein assemblies with tunable phase diagrams in living cells," Nat. Chem. Biol. **16**, 939 (2020).

[41] V. N. Uversky, "Intrinsically disordered proteins in overcrowded milieu: Membrane-less organelles, phase separation, and intrinsic disorder," Curr. Opin. Struct. Biol. **44**, 18 (2017).

[42] W. M. Aumiller and C. D. Keating, "Experimental models for dynamic compartmentalization of biomolecules in liquid organelles: Reversible formation and partitioning in aqueous biphasic systems," Adv. Colloid Interface Sci. **239**, 75 (2017).

[43] Y.-H. Lin, J. Song, J. D. Forman-Kay, and H. S. Chan, "Random-phase-approximation theory for sequence-dependent, biologically functional liquid-liquid phase separation of intrinsically disordered proteins," J. Mol. Liq. **228**, 176 (2017).

[44] Y.-H. Lin and H. S. Chan, "Phase separation and single-chain compactness of charged disordered proteins are strongly correlated," Biophys. J. **112**, 2043 (2017).

[45] S. Das, A. Amin, Y.-H. Lin, and H. Chan, "Coarse-grained residue-based models of disordered protein condensates: Utility and limitations of simple charge pattern parameters," Phys. Chem. Chem. Phys. **20**, 28558 (2018).

[46] M. Paloni, R. Bailly, L. Ciandrini, and A. Barducci, "Unraveling molecular interactions in liquid-liquid phase-separation of disordered proteins by atomistic simulations," J. Phys. Chem. B **124**, 9009 (2020).

[47] T. J. Welsh, G. Krainer, J. R. Espinosa, J. A. Joseph, A. Sridhar, M. Jahnel, W. E. Arter, K. L. Saar, S. Alberti, R. Collepardo-Guevara, and T. P. Knowles, "Surface electrostatics govern the emulsion stability of biomolecular condensates," bioRxiv (2020), 10.1101/2020.04.20.047910.

[48] W. Zheng, G. L. Dignon, N. Jovic, X. Xu, R. M. Regy, N. L. Fawzi, Y. C. Kim, R. B. Best, and J. Mittal, "Molecular details of protein condensates probed by microsecond long atomistic simulations," J. Phys. Chem. B **124**, 11671 (2020).

[49] A. Vitalis and R. V. Pappu, "ABSINTH: A new continuum solvation model for simulations of polypeptides in aqueous solutions," J. Comput. Chem. **30**, 673 (2009).

[50] A. Vitalis and R. V. Pappu, "Methods for Monte Carlo simulations of biomacromolecules," Annu. Rep. Comput. Chem. **5**, 49 (2009).

[51] X. Zeng, A. S. Holehouse, A. Chilkoti, T. Mittag, and R. V. Pappu, "Connecting coil-to-globule transitions to full phase diagrams for intrinsically disordered proteins," Biophys. J. **119**, 402 (2020).

[52] V. Nguemaha and H.-X. Zhou, "Liquid-liquid phase separation of patchy particles illuminates diverse effects of regulatory components on protein droplet formation," Sci. Rep. **8**, 6728 (2018).

[53] A. Reinhardt, A. J. Williamson, J. P. K. Doye, J. Carrete, L. M. Varela, and A. A. Louis, "Re-entrant phase behavior for systems with competition between phase separation and self-assembly," J. Chem. Phys. **134**, 104905 (2011).

[54] H. Liu, S. K. Kumar, and F. Sciortino, "Vapor-liquid coexistence of patchy models: Relevance to protein phase behavior," J. Chem. Phys. **127**, 084902 (2007).

[55] T. S. Harmon, A. S. Holehouse, M. K. Rosen, and R. V. Pappu, "Intrinsically disordered linkers determine the interplay between phase separation and gelation in multivalent proteins," eLife **6**, e30294 (2017).

[56] T. S. Harmon, A. S. Holehouse, and R. V. Pappu, "Differential solvation of intrinsically disordered linkers drives the formation of spatially organized droplets in ternary systems of linear multivalent proteins," New J. Phys. **20**, 045002 (2018).

[57] K. M. Ruff, T. S. Harmon, and R. V. Pappu, "CAMELOT: A machine learning approach for coarse-grained simulations of aggregation of block-copolymeric protein sequences," J. Chem. Phys. **143**, 243123 (2015).

[58] S. Das, Y.-H. Lin, R. M. Vernon, J. D. Forman-Kay, and H. S. Chan, "Comparative roles of charge, π, and hydrophobic interactions in sequence-dependent phase separation of intrinsically disordered proteins," Proc. Natl. Acad. Sci. U. S. A. **117**, 28795 (2020).

[59] R. M. Regy, G. L. Dignon, W. Zheng, Y. C. Kim, and J. Mittal, "Sequence dependent phase separation of protein-polynucleotide mixtures elucidated using molecular simulations," Nucleic Acids Res. **48**, 12593 (2020).

[60] R. S. Judson and H. Rabitz, "Teaching lasers to control molecules," Phys. Rev. Lett. **68**, 1500 (1992).

[61] A. Assion, T. Baumert, M. Bergt, T. Brixner, B. Kiefer, V. Seyfried, M. Strehle, and G. Gerber, "Control of chemical reactions by feedback-optimized phase-shaped femtosecond laser pulses," Science **282**, 919 (1998).

[62] D. M. Deaven and K. M. Ho, "Molecular geometry optimization with a genetic algorithm," Phys. Rev. Lett. **75**, 288 (1995).

[63] S. M. Woodley, P. D. Battle, J. D. Gale, and C. R. A. Catlow, "The prediction of inorganic crystal structures using a genetic algorithm and energy minimisation," Phys. Chem. Chem. Phys. **1**, 2535 (1999).

[64] T. Dandekar and P. Argos, "Potential of genetic algorithms in protein folding and protein engineering simulations," Protein Eng. Des. Sel. **5**, 637 (1992).

[65] R. Unger and J. Moult, "Genetic algorithms for protein folding simulations," J. Molec. Biol. **231**, 75 (1993).

[66] W. P. C. Stemmer, "Rapid evolution of a protein in vitro by DNA shuffling," Nature **370**, 389 (1994).

[67] J. O. Spiegel and J. D. Durrant, "AutoGrow4: an open-source genetic algorithm for de novo drug design and lead optimization," J. Cheminformatics **12** (2020), 10.1186/s13321-020-00429-4.

[68] G. J. Pauschenwein and G. Kahl, "Clusters, columns, and lamellae—minimum energy configurations in core softened potentials," Soft Matter **4**, 1396 (2008).

[69] J. Fornleitner and G. Kahl, "Lane formation vs. cluster formation in two-dimensional square-shoulder systems — A genetic algorithm approach," EPL **82**, 18001 (2008).

[70] L. Filion and M. Dijkstra, "Prediction of binary hard-sphere crystal structures," Phys. Rev. E **79**, 046714 (2009).

[71] I. G. Johnston, S. E. Ahnert, J. P. K. Doye, and A. A. Louis, "Evolutionary dynamics in a simple model of self-assembly," Phys. Rev. E **83**, 066105 (2011).

[72] M. Z. Miskin and H. M. Jaeger, "Adapting granular materials through artificial evolution," Nat. Mater. **12**, 326 (2013).

[73] J. C. Forster, J. Krausser, M. R. Vuyyuru, B. Baum, and A. Šarić, "Exploring the design rules for efficient membrane-reshaping nanostructures," Phys. Rev. Lett. **125**, 228101 (2020).

[74] X. Zeng, C. Liu, M. J. Fossat, P. Ren, A. Chilkoti, and R. V. Pappu, "Design of intrinsically disordered proteins that undergo phase transitions with lower critical solution temperatures," APL Mater. **9**, 021119 (2021).

[75] C. Vega, E. Sanz, J. L. F. Abascal, and E. G. Noya, "Determination of phase diagrams via computer simulation: methodology and applications to water, electrolytes and proteins," J. Phys.: Condens. Matter **20**, 153101 (2008).

[76] Y. Lin, D. S. Protter, M. K. Rosen, and R. Parker, "Formation and maturation of phase-separated liquid droplets by RNA-binding proteins," Mol. Cell **60**, 208 (2015).

[77] J. Kang, L. Lim, Y. Lu, and J. Song, "A unified mechanism for LLPS of ALS/FTLD-causing FUS as well as its modulation by ATP and oligonucleic acids," PLOS Biology **17**, 1 (2019).

[78] L. H. Kapcha and P. J. Rossky, "A simple atomic-level hydrophobicity scale reveals protein interfacial structure," J. Molec. Biol. **426**, 484 (2014).

[79] Y. Lin, S. L. Currie, and M. K. Rosen, "Intrinsically disordered sequences enable modulation of protein phase separation through distributed tyrosine motifs," J. Biol. Chem. **292**, 19110 (2017).

[80] P. Dasmeh and A. Wagner, "Natural selection on the phase-separation properties of FUS during 160 million years of mammalian evolution," Mol. Biol. Evol. **38**, msaa258 (2021).

[81] M. Kato, T. W. Han, S. Xie, K. Shi, X. Du, L. C. Wu, H. Mirzaei, E. J. Goldsmith, J. Longgood, J. Pei, N. V. Grishin, D. E. Frantz, J. W. Schneider, S. Chen, L. Li, M. R. Sawaya, D. Eisenberg, R. Tycko, and S. L. McKnight, "Cell-free formation of RNA granules: Low complexity sequence domains form dynamic fibers within hydrogels," Cell **149**, 753 (2012).

[82] Z. Monahan, V. H. Ryan, A. M. Janke, K. A. Burke, S. N. Rhoads, G. H. Zerze, R. O'Meally, G. L. Dignon, A. E. Conicella, W. Zheng, R. B. Best, R. N. Cole, J. Mittal, F. Shewmaker, and N. L. Fawzi, "Phosphorylation of the FUS low-complexity domain disrupts phase separation, aggregation, and toxicity," EMBO J. **36**, 2951 (2017).

[83] K. L. Morrison and G. A. Weiss, "Combinatorial alanine-scanning," Curr. Op. Chem. Biol. **5**, 302 (2001).

[84] B. Cunningham and J. Wells, "High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis," Science **244**, 1081 (1989).

[85] T. Kortemme, D. E. Kim, and D. Baker, "Computational alanine scanning of protein-protein interfaces," Sci. Signaling **2004**, pl2 (2004).

[86] K. A. Scott, D. O. V. Alonso, S. Sato, A. R. Fersht, and V. Daggett, "Conformational entropy of alanine versus glycine in protein denatured states," Proc. Natl. Acad. Sci. U. S. A. **104**, 2661 (2007).

[87] L. Sawle and K. Ghosh, "A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins," J. Chem. Phys. **143**, 085101 (2015).

[88] C. W. Pak, M. Kosno, A. S. Holehouse, S. B. Padrick, A. Mittal, R. Ali, A. A. Yunus, D. R. Liu, R. V. Pappu, and M. K. Rosen, "Sequence determinants of intracellular phase separation by complex coacervation of a disordered protein," Mol. Cell **63**, 72 (2016).

[89] T. J. Nott, E. Petsalaki, P. Farber, D. Jervis, E. Fussner, A. Plochowietz, T. D. Craggs, D. P. Bazett-Jones, T. Pawson, J. D. Forman-Kay, and A. J. Baldwin, "Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles," Mol. Cell **57**, 936 (2015).

[90] Y.-H. Lin, J. P. Brady, J. D. Forman-Kay, and H. S. Chan, "Charge pattern matching as a 'fuzzy' mode of molecular recognition for the functional phase separations of intrinsically disordered proteins," New J. Phys. **19**, 115003 (2017).

[91] J. McCarty, K. T. Delaney, S. P. O. Danielsen, G. H. Fredrickson, and J.-E. Shea, "Complete phase diagram for liquid–liquid phase separation of intrinsically disordered proteins," J. Phys. Chem. Lett. **10**, 1644 (2019).

[92] S. Elbaum-Garfinkle, Y. Kim, K. Szczepaniak, C. C.-H. Chen, C. R. Eckmann, S. Myong, and C. P. Brangwynne, "The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics," Proc. Natl. Acad. Sci. U. S. A. **112**, 7189 (2015).

[93] J. A. Joseph, J. R. Espinosa, I. Sanchez-Burgos, A. Garaizar, D. Frenkel, and R. Collepardo-Guevara, "Thermodynamics and kinetics of phase separation of protein-RNA mixtures by a minimal model," Biophys. J. **120**, 1219 (2021).

[94] I. Sanchez-Burgos, J. R. Espinosa, J. A. Joseph, and R. Collepardo-Guevara, "Valency and binding affinity variations can regulate the multilayered organization of protein condensates with many components," Biomolecules **11**, 278 (2021).

[95] I. Sanchez-Burgos, J. A. Joseph, R. Collepardo-Guevara, and J. R. Espinosa, "Size conservation emerges spontaneously in biomolecular condensates formed by scaffolds and surfactant clients," Sci. Rep. **11** (2021), 10.1038/s41598-021-94309-y.

[96] H. Zhang, C. Li, F. Yang, J. Su, J. Tan, X. Zhang, and C. Wang, "Cation-pi interactions at non-redundant protein-RNA interfaces," Biochemistry (Moscow) **79**, 643 (2014).

[97] P. R. Banerjee, A. N. Milin, M. M. Moosa, P. L. Onuchic, and A. A. Deniz, "Reentrant phase transition drives dynamic substructure formation in ribonucleoprotein droplets," Angew. Chem., Int. Ed. **56**, 11354 (2017).

[98] I. Alshareedah, T. Kaur, J. Ngo, H. Seppala, L.-A. D. Kounatse, W. Wang, M. M. Moosa, and P. R. Banerjee, "Interplay between short-range attraction and long-range repulsion controls reentrant liquid condensation of ribonucleoprotein–RNA complexes," J. Am. Chem. Soc. **141**, 14593 (2019).

[99] G. Raos and G. Allegra, "Chain collapse and phase separation in poor-solvent polymer solutions: A unified molecular description," J. Chem. Phys. **104**, 1626 (1996).

[100] A. H. Mao, S. L. Crick, A. Vitalis, C. L. Chicoine, and R. V. Pappu, "Net charge per residue modulates conformational ensembles of intrinsically disordered proteins," Proc. Natl. Acad. Sci. U. S. A. **107**, 8183 (2010).

[101] R. K. Das, Y. Huang, A. H. Phillips, R. W. Kriwacki, and R. V. Pappu, "Cryptic sequence features within the disordered protein p27Kip1 regulate cell cycle signaling," Proc. Natl. Acad. Sci. U. S. A. **113**, 5616 (2016).

[102] E. W. Martin, A. S. Holehouse, C. R. Grace, A. Hughes, R. V. Pappu, and T. Mittag, "Sequence determinants of the conformational properties of an intrinsically disordered protein prior to and upon multisite phosphorylation," J. Am. Chem. Soc. **138**, 15323 (2016).

[103] K. P. Sherry, R. K. Das, R. V. Pappu, and D. Barrick, "Control of transcriptional activity by design of charge patterning in the intrinsically disordered RAM region of the notch receptor," Proc. Natl. Acad. Sci. U. S. A. **114**, E9243 (2017).

[104] A. E. Conicella, G. L. Dignon, G. H. Zerze, H. B. Schmidt, A. M. D'Ordine, Y. C. Kim, R. Rohatgi, Y. M. Ayala, J. Mittal, and N. L. Fawzi, "TDP-43 α-helical structure tunes liquid–liquid phase separation and function," Proc. Natl. Acad. Sci. U. S. A. **117**, 5883 (2020).

[105] T. M. Perdikari, N. Jovic, G. L. Dignon, Y. C. Kim, N. L. Fawzi, and J. Mittal, "A predictive coarse-grained model for position-specific effects of post-translational modifications," Biophys. J. **120**, 1187 (2021).

[106] K. You, Q. Huang, C. Yu, B. Shen, C. Sevilla, M. Shi, H. Hermjakob, Y. Chen, and T. Li, "PhaSepDB: a database of liquid–liquid phase separation related proteins," Nucleic Acids Res. **48**, D354 (2020).

[107] B. Mészáros, G. Erdős, B. Szabó, É. Schád, Á. Tantos, R. Abukhairan, T. Horváth, N. Murvai, O. P. Kovács, M. Kovács, S. C. E. Tosatto, P. Tompa, Z. Dosztányi, and R. Pancsa, "PhaSePro: the database of proteins driving liquid–liquid phase separation," Nucleic Acids Res. **48**, D360 (2020).

[108] D. Piovesan, M. Necci, N. Escobedo, A. M. Monzon, A. Hatos, I. Mičetić, F. Quaglia, L. Paladin, P. Ramasamy, Z. Dosztányi, W. F. Vranken, N. E. Davey, G. Parisi, M. Fuxreiter, and S. C. E. Tosatto, "MobiDB: intrinsically disordered proteins in 2021," Nucleic Acids Res. **49**, D361 (2021).

[109] L. Chambers, *Practical handbook of genetic algorithms: applications*, Vol. 1 (CRC Press Inc., 1995).

[110] M. Srinivas and L. M. Patnaik, "Genetic algorithms: a survey," Computer **27**, 17 (1994).

[111] B. L. Miller and D. E. Goldberg, "Genetic algorithms, tournament selection, and the effects of noise," Complex Systems **9**, 193 (1995).

[112] E. Cantú-Paz, *Efficient and Accurate Parallel Genetic Algorithms* (Springer, 2000).

[113] H. S. Ashbaugh and H. W. Hatch, "Natively unfolded protein stability as a coil-to-globule transition in charge/hydropathy space," J. Am. Chem. Soc. **130**, 9536 (2008).

[114] P. Debye and E. Hückel, "Zur Theorie der Elektrolyte. I. Gefrierpunktserniedrigung und verwandte Erscheinungen," Phys. Z. **24**, 185 (1923).

[115] Y. C. Kim and G. Hummer, "Coarse-grained models for simulations of multiprotein complexes: Application to ubiquitin binding," J. Mol. Biol. **375**, 1416 (2008).

[116] S. Miyazawa and R. L. Jernigan, "Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading," J. Mol. Biol. **256**, 623 (1996).

[117] A. Opitz, "Molecular dynamics investigation of a free surface of liquid argon," Phys. Lett. A **47**, 439 (1974).
[118] G. L. Dignon, W. Zheng, and J. Mittal, "Simulation methods for liquid–liquid phase separation of disordered proteins," Curr. Opin. Chem. Eng. **23**, 92 (2019).
[119] S. Plimpton, "Fast parallel algorithms for short-range molecular dynamics," J. Comput. Phys. **117**, 1 (1995); P. in 't Veld, S. Plimpton, and G. Grest, "Accurate and efficient methods for modeling colloidal mixtures in an explicit solvent using molecular dynamics," Comp. Phys. Commun. **179**, 320 (2008).
[120] T. Schneider and E. Stoll, "Molecular-dynamics study of a three-dimensional one-component model for distortive phase transitions," Phys. Rev. B **17**, 1302 (1978).
[121] D. Vitkup, D. Ringe, G. A. Petsko, and M. Karplus, "Solvent mobility and the protein 'glass' transition," Nat. Struct. Biol. **7**, 34 (2000).
[122] I. V. Volgin, S. V. Larin, E. Abad, and S. V. Lyulin, "Molecular dynamics simulations of fullerene diffusion in polymer melts," Macromolecules **50**, 2207 (2017).
[123] B. Kresse, M. Hofmann, A. F. Privalov, N. Fatkullin, F. Fujara, and E. A. Rössler, "All polymer diffusion regimes covered by combining field-cycling and field-gradient ^{1}H NMR," Macromolecules **48**, 4491 (2015).
[124] A. Einstein, "Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen," Ann. Phys.–Leipzig **322**, 549 (1905).